Severity Prediction for COVID-19 Patients via Recurrent Neural Networks

Junghwan Lee
Department of Biomedical Informatics
Columbia University
New York, NY
jl5307@columbia.edu

Casey Ta
Department of Biomedical Informatics
Columbia University
New York, NY
ct2865@cumc.columbia.edu

Jae Hyun Kim
Department of Biomedical Informatics
Columbia University
New York, NY
jk4388@cumc.columbia.edu

Cong Liu*
Department of Biomedical Informatics
Columbia University
New York, NY
cl3720@cumc.columbia.edu

Chunhua Weng*
Department of Biomedical Informatics
Columbia University
New York, NY
cw2384@cumc.columbia.edu

Abstract

The novel coronavirus disease-2019 (COVID-19) pandemic has threatened the health of tens of millions of people worldwide and posed enormous burden on the global healthcare systems. Many prediction models have been proposed to fight against the pandemic. In this paper, we propose a model to predict whether a patient infected with COVID-19 will develop severe outcomes based only on the patient's historical electronic health records (EHR) using recurrent neural networks (RNN). The predicted severity risk score represents the probability for a person to progress into severe status (mechanical ventilation, tracheostomy, or death) after being infected with COVID-19. While many of the existing models use features obtained after diagnosis of COVID-19, our proposed model only utilizes a patient's historical EHR so that it can enable proactive risk management before or at the time of hospital admission.

1 Introduction

The novel coronavirus disease-2019 (COVID-19) has threatened the health of tens of millions of people over the world and imposed an enormous burden on global healthcare systems. To fight against the pandemic and mitigate the burden, numerous efforts have been made by scientists for developing risk prediction models for COVID-19. Prognostic models, among the most important risk prediction models, have been developed to predict risks of mortality [27, 18, 12] and progression to severe status [16, 11, 5] for COVID-19 patients. Commonly used predictors for those prognostic models for COVID-19 patients include comorbidities, age, sex, lab test results (e.g., lymphocyte count, C reactive protein, creatinine), and radiologic imaging features [26]. The existing models, however, spanning from Cox proportional hazards models to state-of-the-art machine learning or deep learning models, heavily rely on features obtained after hospital admission or diagnosis of COVID-19

Preprint. Under review.

Table 1: Characteristics of the COVID-19 cohort. Senior patients: aged ≥ 65 at hospital admission related to COVID-19. SD: standard deviation.

Total # of patients	2,498
# of severe patients (%)	569 (22.8%)
# of senior patients (%)	1,389 (55.6%)
# of male patients (%)	1,209 (48.4%)
Unique # of condition concepts appeared	6,280
Avg. # of visits per patient (SD)	71.78 (114.83)
Avg. # of concepts per visit (SD)	2.61 (3.08)

for post-diagnosis prognosis [26]. A major limitation of these methods is that they apply only to people with the clinical features needed by the models after the diagnosis of COVID-19.

Recurrent neural networks (RNN) have been widely used in modeling sequential data such as speech and language due to its strengths of capturing hidden relationships between the sequential inputs [15]. There have been several studies in the healthcare domain that used RNNs to predict future medical events or risk scores of certain diseases, leveraging the sequential nature of electronic health records (EHR). Lipton et al., [17] and Choi et al., [7] both used RNNs for predicting future medical events based on the historical EHR data. Choi et al., [8] published related work using RNN for predicting the risk score of heart failure based on the patient's historical medical records.

In this study, we applied an RNN on a patient's historical EHR to predict the patient's risk of severe outcomes from COVID-19, representing the probability for a person to progress into severe status after being infected with COVID-19. One major advantage of this method is that it does not require any data after the diagnosis of COVID-19, so that it can predict the risk of severe outcomes of the people regardless of the infection status, allowing proactive risk management and resource allocation in advance, which can be critical for health policy makers and hospital administrators.

2 Methods

2.1 COVID-19 Cohort Description

New York City was one of the epicenters of the COVID-19 pandemic in March and April of 2020. NewYork Presbyterian Hospital/Columbia University Irving Medical Center (NYP/CUIMC) treated a large cohort of COVID-19 patients. For this work, we obtained longitudinal EHR data from NYP/CUIMC's Observational Medical Outcomes Partnership (OMOP) database, which contains 30 years' worth of comprehensive EHR data for about 6.5 million patients. CUIMC transitioned EHR systems to Epic on February 1, 2020. Therefore, we also obtained all EHR data from Epic and Clarity EHR systems for the patients infected with COVID-19 in NYP/CUIMC updated until May 31, 2020. These data were transformed into the OMOP Common Data Model in June 2020. For this study, we only used condition (i.e. diagnosis) concepts as predictors for the models. This study received institutional review board approval (AAAR3954) with a waiver for informed consent.

The COVID-19 cohort was identified as patients 18 years or older who were hospitalized and tested positive for SARS-CoV-2 within 21 days before or during their hospitalization. Patients also must have at least one visit record prior to March 1, 2020. We obtained all historical condition concepts and demographic information prior to the hospital admission due to infection of COVID-19 for the identified patients in the cohort in longitudinal order. Basic characteristics of the COVID-19 cohort are shown in Table 1. The hospitalized COVID-19 patients were classified into two groups: severe vs. moderate. Severe COVID-19 patients were identified as patients who had at least one of the following events during hospitalization: mechanical ventilation, tracheostomy, or death; these events correspond to a severity score of ≥ 6 in the World Health Organization ordinal scale for clinical improvement [25]. Moderate COVID-19 patients refer to the patients who were either discharged from hospitalization or were still hospitalized but without any signal of the three severe events.

2.2 Problem Definition

For each person in the cohort, all historical inpatient or outpatient visits were extracted in the form of multi-hot encoding x_i for $i = 1, ..., n$, where n is the number of total visits that the person made before the hospital admission related to COVID-19. The multi-hot encoding $x_i \in [0, 1]^k$ represents the i-th visit of the person, where k denotes the number of unique medical concepts appeared in the cohort and x_i^l is 1 if the l-th medical concept appears in the patient's i-th visit. Our goal is to predict the patient's risk of developing severe outcomes based on the patient's historical EHR data. The predicted severity risk score ranges between 0 and 1 and represents the estimated probability of the person developing a severe outcome from COVID-19.

2.3 Gated Recurrent Units Preliminaries

We used Gated Recurrent Units [6] (GRU) for the RNN model in this work. Although Long Short Term Memory [10] (LSTM) is the most widely used recurrent unit among all other variants and generally outperforms GRUs on big datasets [3, 23], GRUs show comparable or better performance on tasks with relatively small datasets since GRUs have fewer parameters to train [9]. Figure 1a depicts the architecture of the GRU cell, where x_i, z_i, r_i, $\tilde{h}_i$, and h_i denote the input, update gate, reset gate, candidate hidden state, and hidden state, respectively, all at timestamp t_i. W_h, W_z, W_r, U_h, U_z, and U_r are the trainable weight matrices. The mathematical formulation of the GRU cell is provided in Eq(1) through Eq(4):

$$z_i = \sigma(W_z x_i + U_z h_{i-1} + b_z) \tag{1}$$

$$r_i = \sigma(W_r x_i + U_r h_{i-1} + b_r) \tag{2}$$

$$\tilde{h}_i = tanh(W_h x_i + r_i \circ U_h h_{i-1} + b_h) \tag{3}$$

$$h_i = z_i \circ h_{i-1} + (1 - z_i) \circ \tilde{h}_i \tag{4}$$

where $\circ$ denotes element-wise multiplication, $tanh()$ denotes the hyperbolic tangent function, and $\sigma()$ denotes the sigmoid function. The update gate z_i decides how much information should be updated from the input and is computed as Eq(1). Similarly, the reset gate r_i decides how much information should be ignored from the past information and is computed as Eq(2). The candidate hidden state $\tilde{h}_i$ is computed as Eq(3) using the input and the hidden state of the previous timestamp h_{i-1}. Finally, the hidden state at timestamp t_i is computed as Eq(4), using the candidate hidden state $\tilde{h}_i$ and the hidden state of the previous timestamp h_{i-1}. Since the reset and update gates decide how information from the history of the past inputs should be combined with current inputs to form the new the hidden state, we can utilize the hidden state as the vector containing the information about the patient's total medical history.

2.4 Model Architecture

In order to predict the COVID-19 severity risk score, we proposed an RNN model as depicted in Figure 1b. The model receives as input the patient's visits x_i and outputs the hidden state h_i at each timestamp t_i. For efficient training of the model, we used an embedding layer that transforms the multi-hot encoded visit input x_i into a lower-dimensional embedding (described below). The hidden state at the last timestamp is concatenated with the patient's demographic information vector, and then a single dense layer and logistic regression are sequentially applied to generate the risk score of the patient as defined in Eq(5) and Eq(6):

$$o_{dense} = tanh(W_{dense}[h_T, d] + b_{dense}) \tag{5}$$

$$\hat{y} = \sigma(W_{lr} o_{dense} + b_{lr}) \tag{6}$$

where W_{dense}, W_{lr}, b_{dense}, b_{lr}, o_{dense}, h_T, d and $\hat{y}$ denote the weight matrix of the dense layer, weight matrix of the logistic regression, bias vector of the dense layer, bias of the logistic regression, the output vector of the dense layer, the hidden state at the last timestamp, demographic information vector, and predicted risk score of the patient, respectively. $[,]$ denotes vector concatenation. The patient demographic information vector is a simple concatenation of sex and age of the patient. Sex

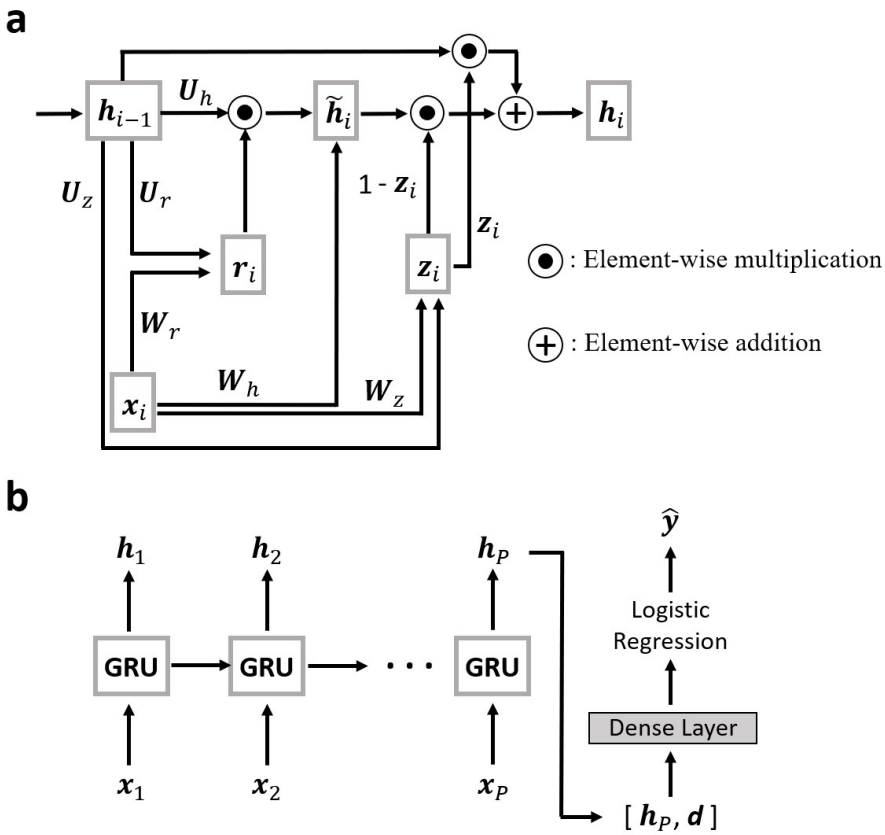

Figure 1: The architecture of (a) the GRU cell and (b) the proposed RNN model.

information is represented as one-hot encoding (i.e. [1, 0] for male and [0, 1] for female), and age is normalized using min-max normalization.

The true label y for each patient was determined based on the outcome status of the patient as observed in the CUIMC database; we assigned 1 for severe patients and 0 for moderate patients. Since the severe and moderate cases were imbalanced, we used weighted cross entropy to calculate the loss, defined as Eq(7):

$$L = -\sum_{j=1}^{N}(wy^{(j)} \log \hat{y}^{(j)} + (1-w)(1-y^{(j)}) \log (1-\hat{y}^{(j)})) \tag{7}$$

where $y^{(j)}$, $\hat{y}^{(j)}$, N, and w are the true label for the j-th patient, the predicted risk score for the j-th patient, the total number of patients in the dataset, and weight for the cross entropy. We used 0.75 for the weight of the cross entropy based on the ratio of the severe and moderate patients in the data to provide more weight on accurately predicting severe cases. Optimization was performed using mini-batch scheme.

3 Results

3.1 Experiment Setup

To evaluate the performance of the RNN model, we compared the average area under the receiver operating characteristic curve (AUC) based on 5-fold cross validation with two other baselines – logistic regression and multilayer perceptron (MLP). The entire dataset was divided into 5 chunks: 3,

1, and 1 chunk(s) were allocated to the training set, validation set, and test set respectively. Different combinations of chunks were allocated to the training set, validation set, and test set at every fold, thus the model is trained, validated, and tested on different datasets at every fold. All models were trained with a maximum of 20 epochs at every fold, and the best model on the validation set was evaluated on the test set. We report the mean and standard deviation of the AUC of all 5 folds based on the test set.

We experimented with two different initializations of the embedding layer in the RNN model: (1) the embedding layer initialized with a random normal distribution; (2) the embedding layer initialized with pre-trained embeddings. Random normal distribution with mean 0 and standard deviation 0.01 was chosen for initialization since it showed better performance than many other baselines in word embedding tasks [14]. We obtained pre-trained embedding using GloVe [21] on the entire dataset of the cohort for 30 epochs. Pre-trained embedding captures the relationships between the medical concepts since GloVe uses the global co-occurrence matrix of concepts for its training, where the co-occurrence matrix is calculated based on the concept co-occurrence in every visit of the patients. The embedding layer was fine-tuned jointly with the prediction task of the model. The dimension of all embeddings was set to 128 since it empirically showed the best performance.

Since mini-batch training shows good generalization performance when the data are relatively small [19], we used a small batch of size 2 in the training. We also empirically found that prediction performance of the model decreased with larger batch sizes. To prevent the model from overfitting, we used dropout [22] with dropout rate 0.3 and norm-2 regularization with regularization coefficient 0.001. The dropout layer was applied between the RNN layer and the dense layer of the RNN model, and norm-2 regularization was applied to all weights in the dense layer and the logistic regression.

3.2 Baselines

3.2.1 Logistic regression

A simple logistic regression model was used for the baseline with three different types of input: aggregated one-hot vector, aggregated embedding and aggregated pre-trained embedding. For each patient, aggregated one-hot vector is the summation of all inputs x_i at all timestamps (corresponding to each visit), which can be understood as the aggregation of all historical records of the patient. In the aggregated embedding and aggregated pre-trained embedding baseline models, the aggregated one-hot vector was passed through an embedding layer which was either initialized using random normal distribution or embeddings pre-trained with GloVe, respectively, using the same initialization schemes as with the RNN model. The dimensions of all embeddings were set to 128. All aggregated inputs were normalized to zero mean and unit variance. The norm-2 regularization with regularization coefficient 0.001 was applied to all weights in the model to reduce overfitting.

3.2.2 Multilayer perceptron

MLP with a single hidden layer was used for another baseline. A dense layer with hyperbolic tangent activation was used for the hidden layer and was followed by the output layer for logistic regression. The three different types of inputs (aggregated one-hot vector, aggregated embedding and aggregated pre-trained embedding) were used with the same settings as described in the subsection above. The number of hidden units in the MLP for aggregated one-hot vector was set to 1,000, and the number of hidden units in the MLPs for the embedded inputs was set to 100. Norm-2 regularization with regularization coefficient 0.001 was applied to all weights to reduce overfitting.

3.3 Implementation Details

We used Tensorflow [1] 2.0.0 to implement the RNN model and all baselines. Adam [13] was used for optimization in training for all models. A machine equipped with 2 $\times$ Intel Xeon Silver 4110 CPUs and 188GB RAM was used. The source codes to implement all models are publicly available at https://github.com/Jayaos/rnn-covid.

3.4 Risk Prediction Performance

We calculated the average AUC of 5-fold cross validation to evaluate the prediction performance of all models (Figure 2). Overall, the RNN model with pretrained embedding achieved the highest

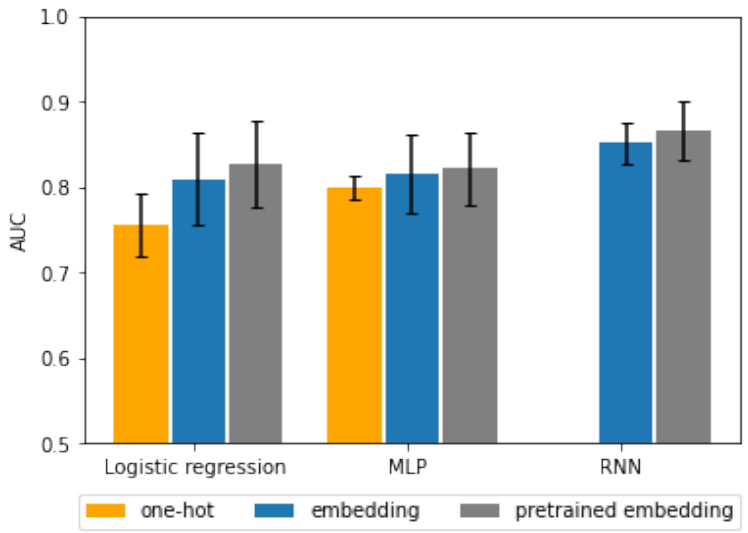

Figure 2: Average 5-fold cross validation AUC of all models.

Table 2: Prediction time per patient for all models.

	Logistic regression	MLP	RNN
Prediction time per patient (seconds)	0.000025	0.000027	0.028613

average AUC (0.865). The RNN model outperformed all other baselines in two different initialization schemes.

3.5 Prediction Time

Table 2 shows the time required to make a prediction per patient for each model. We used the models initialized with pre-trained embedding to measure the prediction time. Other setups, including the machine, were the same as previously described for the training step.

3.6 Analysis of the Predicted Severity Risk Score

We analyzed the predicted severity risk scores of the patients against basic characteristics of the patients to understand how the predicted risk score is affected by patient characteristics and to validate the predicted risk score. We trained the RNN model using the same scheme as described in the experiment setup section above but only obtained the results of the first fold (i.e. 60% training, 20% validation, and 20% test splits). Figure 3a shows the proportion of the patients progressed into severe status in each of the predicted risk score percentile groups. In Figure 3b, we also compared the ROC curve of the predicted risk score in predicting the outcome status of the patient against using either the normalized age or total historical visit counts to predict severe outcomes. Age (normalized age) and total historical visit counts were selected as comparison baselines since they were expected to serve as proxies of a patient's general health status. Figure 4 shows the scatterplot between the predicted risk score and (a) normalized age, (b) total visit counts, and (c) outcome status. The gray-colored dots represent patients in the test set.

3.7 Visualization of the Patient Vector

The output vector of the dense layer in the RNN model is expected to represent information about the patient's medical history. We visualized the output vectors of patients in 2-dimensional space using uniform manifold approximation and projection [20] (UMAP) to explore the pattern of clusters that can be utilized in patient subtyping. Figure 5 shows the scatterplots of the output vectors of severe

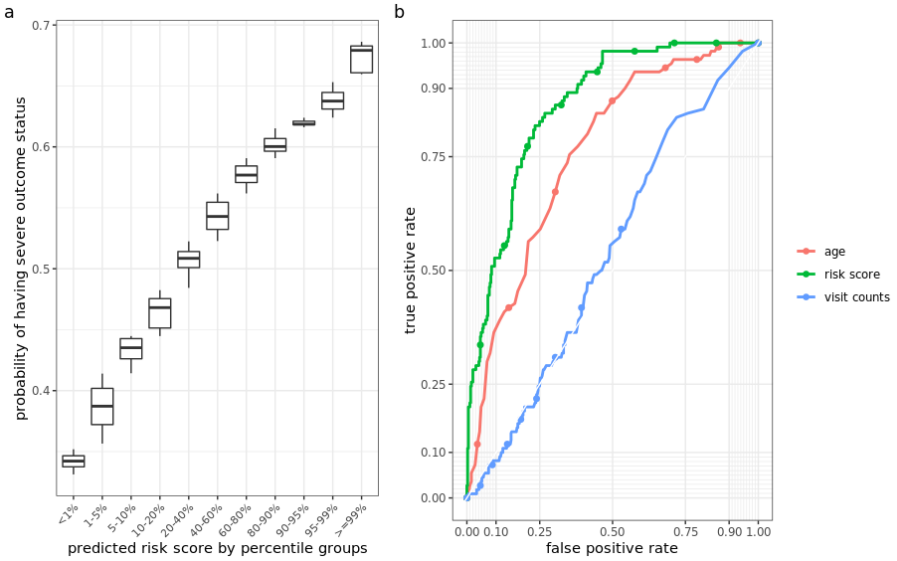

Figure 3: (a) Probability of having severe outcome status in each predicted risk score percentile group (b) Receiver Operating Characteristic (ROC) curve of predicted risk scores, age (normalized age), and total historical visit counts in predicting the outcome status of the patient.

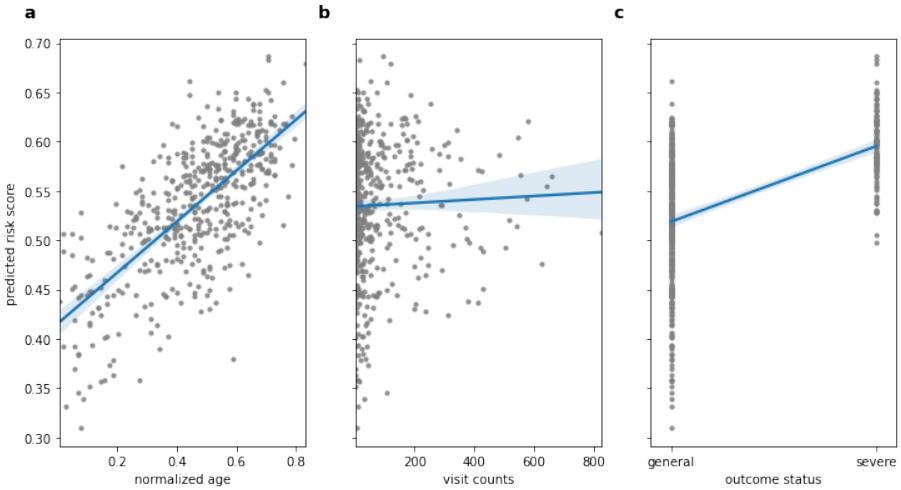

Figure 4: Scatterplot between (a) the normalized age and the predicted risk score, (b) total visit counts and the predicted risk score, and (c) the outcome status and the predicted risk score with the regression line. The gray-colored dots represent patients.

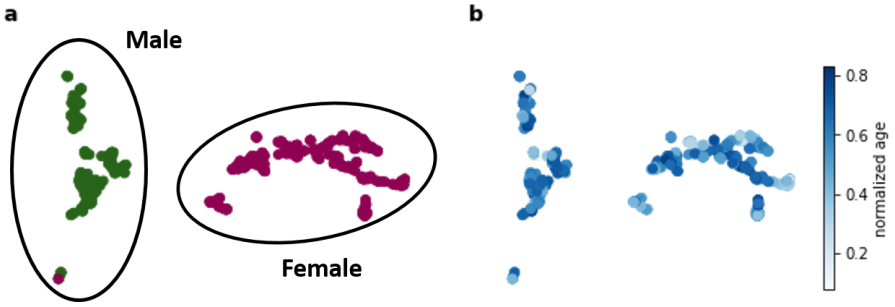

Figure 5: Scatterplots of the output vectors of severe COVID-19 patients. The color of the dots represents sex in a and min-max normalized age in b.

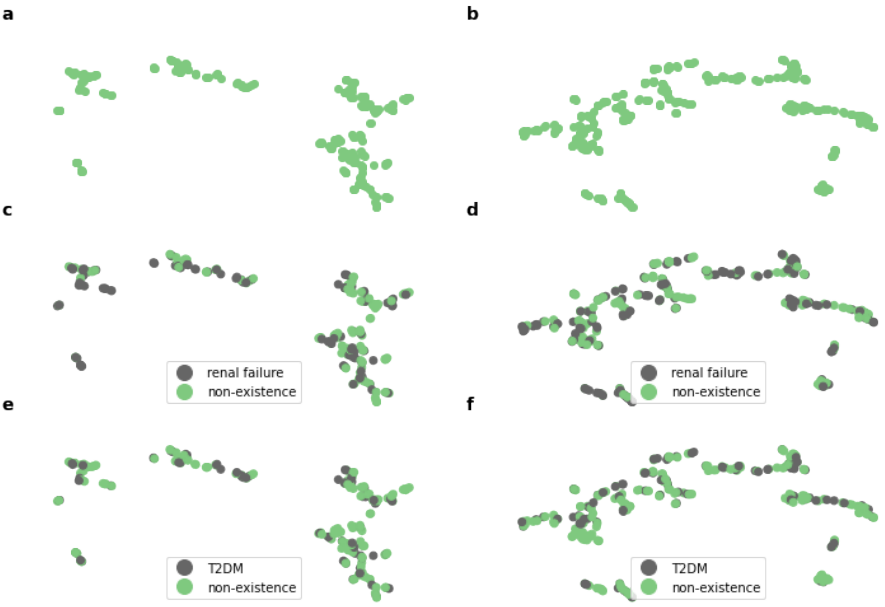

Figure 6: Scatterplots of the output vectors of (a) male and (b) female severe COVID-19 patients. c and d are scatterplots of the output vectors of male and female severe COVID-19 patients with color-labels based on the existence of renal failure. e and f are scatterplots of the output vectors of male and female severe COVID-19 patients with color-labels based on the existence of type 2 diabetes (T2DM).

COVID-19 patients in the entire dataset. The color of the dots represents sex in 5a and min-max normalized age in 5b.

We also visualized the output vectors of male and female severe COVID-19 patients in 2-dimensional space by separately applying UMAP on the patients stratified by sex (Figure 5a and 5b). To further explore characteristics of potential clusters, we color-labeled the dots based on common comorbidities of the cohort. Two common comorbidities of COVID-19 patients in CUIMC, renal failure and type 2 diabetes mellitus (T2DM), were selected [2] (Figure 6c-6f).

4 Discussions

In this study, we proposed an RNN model to predict the severity risk score of COVID-19 patients by utilizing the patients' historical EHR data. The best average AUC (0.865) was achieved by the RNN model initialized with pre-trained embedding. Relatively unstable prediction performances with high variance were observed in the results of the logistic regression models and MLP models compared to the RNN models. This is perhaps because the simple sum of all historical records of patients who

have many or very few historical records might lead to models that cannot be generalized to most of the patients who have an average number of historical records. Looking at the data, the average number of visits for each patient was about 70, while a few patients have more than several hundred visits or less than a few visits. We can see the variance of prediction performance was reduced in the RNN model where the impact of the patients who have many or few visits may have been alleviated by sequential GRU structure.

Although we used a relatively large data set compared to existing COVID-19 studies, which mostly have a few hundred cases [26], the 2,498 cases in our data set is still considered very small for training deep neural network models that contain a large number of parameters to learn. While the model will be able to learn better with more data, obtaining a large data set, however, is not easy for a single institution due to the limited number of patients (and we certainly hope the number of COVID19 patients won't further increase in our institution). We believe obtaining large size of data across different institutions and nations or using other disease cohorts as proxy cohorts will resolve this limitation. One advantage of our approach is that our analysis used a standardized clinical data format, the OMOP Common Data Model. The source code for this analysis can be easily shared with others who have similarly formatted clinical data for evidence aggregation.

While higher accuracies (0.73-0.99) were reported in other studies, the intended use of these models were often not clearly described [26]. The RNN model we propose is intended to aid decision making at the time of or before hospital admission due to COVID-19, hence, only historical EHR data were used in our model. In addition, the RNN model can be applied to the general population to identify people at high risk of developing potential severe outcomes if infected by COVID-19. The RNN model can readily be applied to much larger datasets since it can compute the severity score of all patients in our dataset (about 2,500 patients) within 10 seconds (Table 2).

We validated the effectiveness of the predicted risk score from Figure 3. From Figure 3a, we can see that the patients in higher predicted risk percentiles showed higher probability of having severe outcome status, which indicates the effectiveness of the risk score in predicting severity of patients. Figure 3b shows that the predicted severity risk score obtained from the RNN model can add benefits to the age and total historical visits in predicting severity of patients. Figure 4c shows the predicted severity risk scores of the patients well predict the outcome status of the patients (regression coefficient +0.0765 with p-value < 0.01). We found that there exists a positive relationship between age (normalized age) and the predicted risk score of the patients (regression coefficient +0.2593 with p-value < 0.01). We expected that the number of hospital visits in a patient's medical history would reflect the patient's general health status and that a positive relationship would exist between the number of hospital visits and the predicted risk score. Figure 4b shows, however, that there is no statistically meaningful relationship between the number of hospital visits and the predicted risk score of the patients (p-value 0.4772).

From Figure 5, we can see visible clusters of the patients based on sex and age. Male and female severe patients were divided into two clusters in Figure 5a. Age, however, does not show distinguishable patterns in the clusters from Figure 5b. We can also see clusters within male and female severe patients as shown in Figure 6a and 6b. This suggests that there exist multiple subgroups of severe patients having distinct characteristics within male and female severe patient groups, which shows potential possibility of subtyping COVID-19 patients. These two comorbidities, however, did not show any distinguishable pattern in the clusters (Figure 6c-6f). Thus, we believe that further efforts to uncover detailed characteristics of the clusters will be needed for subtyping COVID-19 patients.

Wynants et al., performed a review of 50 COVID-19 predictive models and reported that most of the models have high risk of bias when evaluated with PROBAST (prediction model risk of bias assessment tool) [26, 24]. They found that two common causes of risk of bias in predictive models for COVID-19 were lack of external validation and selection bias. Our study also has these two limitations. Since the COVID-19 cohort in this work includes patients whose clinical course of care has not yet completed and who may still potentially develop a severe status, there is a chance that discharged patients without any signal of severe status during hospitalization at NYP/CUIMC will later develop a severe status outside of NYP/CUIMC. Future work will include developing an RNN model to predict various status of a patient being infected with COVID-19 rather than simply predicting the risk score. We also plan to modify the RNN model for time-to-event analysis to appropriately handle censored data.

Additionally, the model was not validated with an external cohort. This limitation is mainly caused by medical data exchange issues across different medical institutes, which limits the sharing of medical data across institutions. Since the RNN model is based on a dataset implemented with OMOP common data model, we expect that applying the model to another institution using the common data model will be easily conducted. For example, Burn et al., has performed deep phenotyping on more than 30,000 patients hospitalized with COVID-19 patients in Asian, Europe and American countries using OHDSI network dataset [4]. Future work includes experimenting with and validating the RNN model across different institutions in various countries using the OHDSI network dataset.

5 Conclusion

We proposed a predictive model using recurrent neural networks to predict the severity risk score of COVID-19 patients. The proposed RNN model outperforms logistic regression and multi-layer perceptron models in predicting severe outcome status of COVID-19 patients. We also demonstrated the effectiveness of the predicted severity risk score by analyzing the predicted risk score generated by the RNN model. Future work includes experimenting with the model with a larger dataset and validating the model with an external dataset, as well as further improving the RNN model using more concepts from other domains (e.g., drug, measurements, and procedure) and using time-to-event analysis, which also can address the censored patient issue.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] M. G. Argenziano, S. L. Bruce, C. L. Slater, J. R. Tiao, M. R. Baldwin, R. G. Barr, B. P. Chang, K. H. Chau, J. J. Choi, N. Gavin, et al. Characterization and clinical course of 1000 patients with coronavirus disease 2019 in new york: retrospective case series. *bmj*, 369, 2020.

[3] D. Britz, A. Goldie, M.-T. Luong, and Q. Le. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

[4] E. Burn, S. C. You, A. Sena, K. Kostka, H. Abedtash, M. T. F. Abrahao, A. Alberga, H. Alghoul, O. Alser, T. M. Alshammari, et al. Deep phenotyping of 34,128 patients hospitalised with covid-19 and a comparison with 81,596 influenza patients in america, europe and asia: an international network study. *medRxiv*, 2020.

[5] E. Carr, R. Bendayan, D. Bean, K. O'Gallagher, A. Pickles, D. Stahl, R. Zakeri, T. Searle, A. Shek, Z. Kraljevic, et al. Supplementing the national early warning score (news2) for anticipating early deterioration among patients with covid-19 infection. *medRxiv*, 2020.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[8] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24 (2):361–370, 2017.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[11] H. Huang, S. Cai, Y. Li, Y. Li, Y. Fan, L. Li, C. Lei, X. Tang, F. Hu, F. Li, et al. Prognostic factors for covid-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis. *medRxiv*, 2020.

[12] X. Jianfeng, H. Daniel, C. Hui, T. Abrams Simon, L. Shusheng, W. Guozheng, W. Yishan, K. Hanyujie, B. Laura, Z. Ruiqiang, et al. development and external validation of a prognostic multivariable model on admission for hospitalized patients with covid-19. *https://www. medrxiv. org/content/medrxiv/early/2020/03/30/2020.03. 28.20045997. full. pdf*, 2020.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] T. Kocmi and O. Bojar. An exploration of word embedding initialization in deep-learning tasks. *arXiv preprint arXiv:1711.09160*, 2017.

[15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[16] W. Liang, J. Yao, A. Chen, Q. Lv, M. Zanin, J. Liu, S. Wong, Y. Li, J. Lu, H. Liang, et al. Early triage of critically ill covid-19 patients using deep learning. *Nature communications*, 11(1):1–7, 2020.

[17] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[18] J. Lu, S. Hu, R. Fan, Z. Liu, X. Yin, Q. Wang, Q. Lv, Z. Cai, H. Li, Y. Hu, et al. Acp risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (covid-19) during the early stage of outbreak in wuhan, china. 2020.

[19] D. Masters and C. Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

[20] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

[23] G. Weiss, Y. Goldberg, and E. Yahav. On the practical computational power of finite precision rnns for language recognition. *arXiv preprint arXiv:1805.04908*, 2018.

[24] R. F. Wolff, K. G. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58, 2019.

[25] World Health Organization. Covid-19 therapeutic trial synopsis. URL `https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis`.

[26] L. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, M. C. Haller, G. Heinze, K. G. Moons, R. D. Riley, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj*, 369, 2020.

[27] H. Zhang, T. Shi, X. Wu, X. Zhang, K. Wang, D. Bean, R. Dobson, J. T. Teo, J. Sun, P. Zhao, et al. Risk prediction for poor outcome and death in hospital in-patients with covid-19: derivation in wuhan, china and external validation in london, uk. 2020.