2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-0358-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/BHI50953.2021.9508512

Automated Risk Assessment of COVID-19 Patients at Diagnosis Using Electronic Healthcare Records

Felipe O. Giuste[1], Lawrence L. He[1], Monica Isgut[2], Wenqi Shi[3], Blake J. Anderson[4], and May D. Wang[1]

Abstract—**COVID-19 causes significant morbidity and mortality and early intervention is key to minimizing deadly complications. Available treatments, such as monoclonal antibody therapy, may limit complications, but only when given soon after symptom onset. Unfortunately, these treatments are often expensive, in limited supply, require administration within a hospital setting, and should be given before the onset of severe symptoms. These challenges have created the need for early triage of patients likely to develop life-threatening complications. To meet this need, we developed an automated patient risk assessment model using a real-world hospital system dataset with over 17,000 COVID-positive patients. Specifically, for each COVID-positive patient, we generate a separate risk score for each of four clinical outcomes including death within 30 days, mechanical ventilator use, ICU admission, and any catastrophic event (a superset of dangerous outcomes). We hypothesized that a deep learning binary classification approach can generate these four risk scores from electronic healthcare records data at the time of diagnosis. Our approach achieves significant performance on the four tasks with an area under receiver operating curve (AUROC) for any catastrophic outcome, death within 30 days, ventilator use, and ICU admission of 86.7%, 88.2%, 86.2%, and 87.8%, respectively. In addition, we visualize the sensitivity and specificity of these risk scores to allow clinicians to customize their usage within different clinical outcomes. We believe this work fulfills a clear clinical need for early detection of objective clinical outcomes and can be used for early screening for treatment intervention.**

I. Introduction

Several treatments are available to improve COVID-19 patient outcomes. These include medications that help the body fight the infection early on in the course of the illness, such as monoclonal antibody treatments, preventing high-risk patients from progressing to more severe diseases. It is important to quickly establish a patient risk of severe COVID-19 outcomes to provide an early treatment that maximizes patients' chances of recovering [1], [2]. Thankfully, electronic health records (EHR) provide efficient access to patient data which has created unprecedented opportunities to investigate the properties of clinical events using data-driven approaches [3].

TABLE I
Patient Demographics

Demographics	Patients (%)	Any Catastrophic (%)	Ventilator (%)	ICU (%)	Death30 days (%)
Sex					
Male	7694(43.2)	988(55.5)	562(57.2)	892(55.5)	249(55.5)
Not Male	10111(56.8)	793(44.5)	420(42.8)	715(44.5)	200(44.5)
Race					
African American	7367(41.4)	949(53.3)	565(57.5)	865(53.8)	208(46.3)
Asian	506(2.8)	55(3.1)	29(3)	51(3.2)	13(2.9)
Caucasian	5050(28.4)	576(32.3)	279(28.4)	502(31.2)	183(40.8)
Hispanic	964(5.4)	127(7.1)	67(6.8)	125(7.8)	20(4.5)
Other	3918(22)	74(4.2)	42(4.3)	64(4	25(5.5)
Total	17805	1781	982	1607	449

Clinical decision support systems enabled by artificial intelligence (AI) have been shown to successfully develop patient-specific mortality risk assessments by analyzing available patient clinical data [4]. This work focused on elucidating the clinical features most associated with patient mortality using generalized linear modeling (GLM). They found that past medical history, specifically past history of pneumonia, was an important predictor of mortality. In addition, recent work has also focused on COVID-positive patient risk assessment using EHR early in the disease course. Heldt et al. sought to predict which COVID-positive patients would require mechanical ventilation, had to be admitted to an intensive care unit (ICU), or died due to complications [2]. They used XGboost, a machine learning technique which takes advantage of model boosting and tree-based modeling, to achieve an AUROC of 84%. Many other related studies [5]–[7] of COVID-19 related health outcomes compile demographic statistics offering information focused on single risk factors.

We expand upon existing works by introducing a simple yet effective data-driven approach to asses patient risk for four severe COVID-19 complications. Specifically, we convert the risk assessment task into a binary classification problem to predict whether patients will experience severe COVID-19 complications based on EHR data available during initial COVID-19 diagnosis.

To this end, we generate four separate risk scores for each patient using a dataset of over 17,000 COVID-19 patients and 71 clinical features. These data are obtained from de-identified patients in the Emory University Hospital system. We then implement a deep learning model to generate four negative clinical outcome risk scores for COVID-positive patients. Precisely, we predict the risk of death within 30 days, ICU admission, mechanical ventilation, and a more general

This research was supported by a Wallace H. Coulter Distinguished Faculty Fellowship (M. D. Wang), a Petit Institute Faculty Fellowship (M. D. Wang), and by Microsoft Research.

[1] F. O. Giuste, L. L. He, and M. D. Wang are with the Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: fgiuste@gatech.edu; lhe80@gatech.edu; misgut@gatech.edu; maywang@bme.gatech.edu).

[2] M. Isgut is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: misgut@gatech.edu).

[3] W. Shi is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (email: wqshi@gatech.edu).

[4] B. J. Anderson is with the Department of Medicine, Emory University School of Medicine, Atlanta, GA 30322 USA (email: blake.john.anderson@emory.edu).

978-1-6654-0358-0/21/$31.00 ©2021 IEEE

”any catastrophic outcome” score encompassing all available negative outcome metrics available (death within 90 days, any hospital admission within 90 days, and ventilator use).

We demonstrate that our proposed method outperforms several baseline models and achieves state-of-the-art performance for risk prediction tasks with an area under receiver operating characteristic (AUROC) curve for any catastrophic risk prediction of 86.7% and 88.2% for future mechanical ventilator requirement. We believe this work lays the foundation for rapid patient risk assessment for aggressive intervention and treatment prioritization.

II. Dataset

We used a dataset composed of 17,805 unique patients from the Emory University Hospital system who tested positive for COVID-19 (Table I). A total of 71 clinical features were associated with each patient from the electronic health records from the time of their diagnosis. Specifically, 26 drug categories, 21 vitals and labs, 16 chronic conditions, 6 demographic, and 2 prior hospital visit features (number of prior ER visits and number of prior hospital admissions) were available. In addition, four post-COVID diagnosis binary clinical outcome variables were obtained for use in generating risk scores including death within 30 days (death30), mechanical ventilator use (vent), and ICU admission. In order to generate a more general patient risk score, a binary outcome variable called ”anyCatastrophic” was created and set to 1 if the patient died within 90 days after diagnosis, used a ventilator, was admitted to ICU, or was admitted to hospital within 90 days.

III. Approach

A. Feature Selection

We made as few assumptions as possible about the effects of features on clinical outcomes and allowed all available clinical features at the time of diagnosis to be included in the model if they were not excluded after quality control. This quality control included data cleanup and outlier removal. We removed discrete features with a non-zero value in less than 1% of patients to eliminate the noise that may be introduced in the data by rare, or incorrectly-entered, medications or conditions (i.e. potential outliers). We also removed any features that were missing in more than 75% of patients because these features may not be reliably filled in with data imputation.

After selection, we retained 56 of the 71 original features. These include 17 drug categories, 17 vitals and labs, 14 chronic conditions, 6 demographic, and 2 prior hospital visit features.

B. Data Preprocessing

Patients in the original dataset were randomly assigned into training (60%), validation (20%), or testing (20%) datasets.

To preprocess the data, we normalized the continuous features and used k-Nearest Neighbors (KNN) imputation (K=5) with attention paid to prevent data leakage from the test set into the training and validation sets, and also prevent validation data leakage into the training dataset. Training set preprocessing did not use data from either validation or test sets. For the validation set, we normalized and imputed on the combined training and validation sets to mirror the real-world approach. The testing dataset was preprocessed using data from the full dataset. This strategy reflects the real-world implementation of our model where training data is available during preprocessing and contains patients from the same hospital system to inform normalization and imputation.

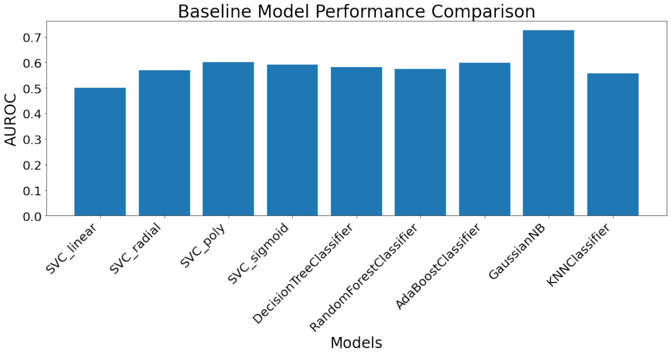


Fig. 1. Histogram of baseline model validation results after training on the training dataset. Gaussian Naive Bayes (NB) achieves an AUROC of 72.6%, outperforming all other baseline classifiers. Results show the best model performance after hyperparameter tuning.

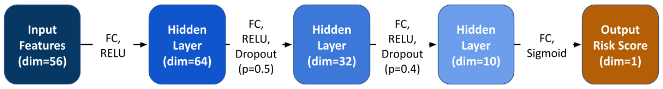


Fig. 2. Optimized risk assessment model architecture. Hyperparameter tuning was conducted on number of model layers, learning rate, and dropout intensity. Four models are trained using this architecture (one for each clinical outcome). Input is 56 clinical features obtained at time of COVID diagnosis from the patient’s electronic health records. Model output is a number between zero and one representing the risk score for a given patient.

C. Baseline Models

To establish baseline classification performance for our four clinical outcomes, we trained nine of the most common machine learning classifiers: support vector machines (SVMs) with four different kernels (linear, radial, polynomial, and sigmoid), random forest, AdaBoost, Gaussian Naïve Bayes, and K-Nearest Neighbors (KNN). We used ‘anyCatastrophic’ as the outcome variable for hyperparameter tuning and performance measurement because it encompasses all negative outcomes relevant for meaningful patient risk assessment. When comparing results for the validation set, we found that Gaussian Naïve Bayes outperformed the others with an area under the receiver operating curve (AUROC) of 72.6%. Figure 1 shows the AUROC for each shallow learning classifier.

D. Deep Learning Classifiers

To improve upon our baseline models, we developed a deep learning model due to its proven track record in clinical decision support [8]. Hyperparameter tuning was conducted on number of hidden layers, dropout intensity, and learning rate. Dropout intensity refers to the probability of setting each

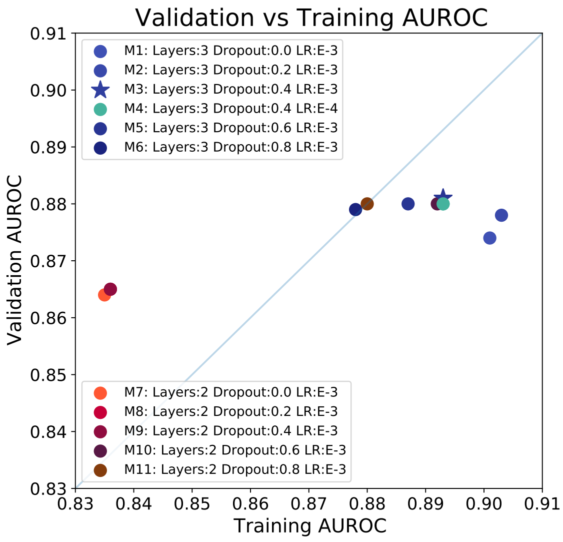


Fig. 3. Validation vs. Training Performance across deep learning model hyperparameters. Models with three hidden layers slightly outperformed those with two. Increasing dropout intensity slightly improved model performance. Learning rate (LR) of 10E-4 was tested for the best performing model, but resulted in slower training times with no significant improvement in validation performance.

input to zero in the first dropout layer. The first dropout layer was always 10% more than the second dropout layer. We used 'anyCatastrophic' performance to compare with baseline models. Parameter optimization was conducted on the training dataset and performance (AUROC) was measured in the validation dataset. To reduce the effects of class imbalances, the label with the most samples was randomly undersampled within each epoch to match the size of the smaller label dataset. Validation versus training performance for each set of hyperparameters was examined to identify the best model for testing (Figure 3). The highest validation AUROC was 88.1% for Model 3. We found the hyperparameter set that yielded the highest validation AUROC had three hidden layers, 0.5 intensity for the first dropout layer, 0.4 intensity for the second dropout layer, and a learning rate of 1E-3 (Figure 2).

Our deep learning classifier outperformed all the baseline classifiers by a considerable amount on the validation data. Therefore, we chose the deep learning classifier on the final testing set. Using the optimal hyperparameters, we trained four deep learning classifiers on the combined training and validation sets, one model for each outcome variable ('anyCatastrophic', ICU admission, ventilator use, and death within 30 days of diagnosis). Final model performance was measured on the unseen test dataset using AUROC, MCC, sensitivity, and specificity (Figure 5).

IV. Results

Figure 4 shows the AUROC for the testing set for each of these outcome variables. Since the model outputs a continuous value between 0 and 1, we need to select a threshold to classify the final output as a binary value. Figure 5 shows the sensitivity, specificity, and Matthews Correlation Coefficient (MCC) for different threshold values for each of the outcome variables. The model yields high AUROC scores for each of the outcome variables, as well as high sensitivity and specificity scores across different thresholds for each of the outcome variables.

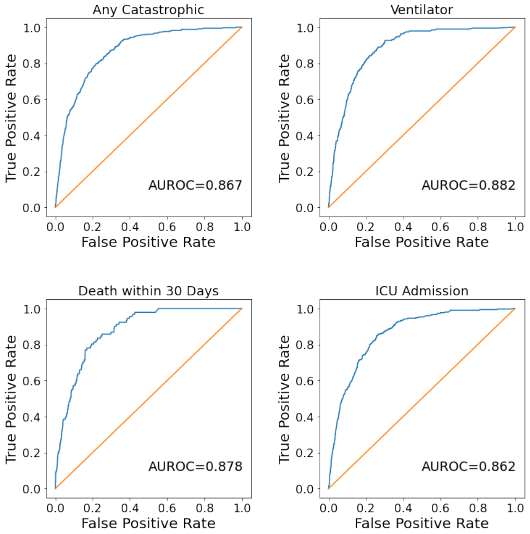
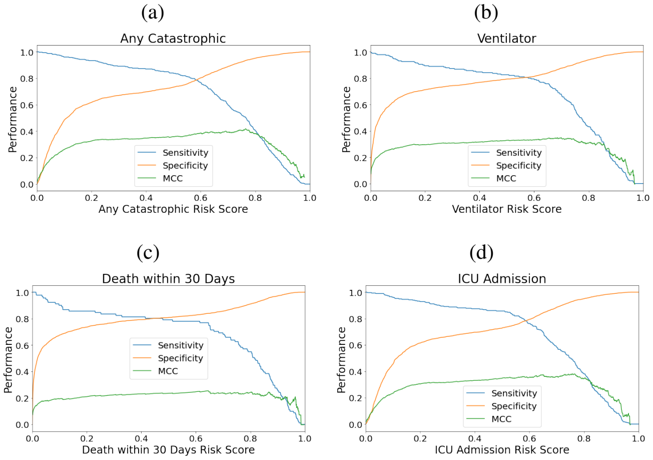


Fig. 4. Area under receiver operating curve (AUROC) of our four models achieves over 85% on the unseen test dataset. Orange lines represent the theoretical results of an uninformative model. Blue lines represent the relationship between true positive rate (TPR) and false positive rate (FPR) of our models.


(a)

(b)

(c)

(d)

Fig. 5. Risk assessment model sensitivity, specificity, and MCC when applied to an unseen test set. As the risk score increases, specificity for the positive class (poor clinical outcome) increases as sensitivity decreases, as expected. a) Prediction of any catastrophic outcome achieves a maximum MCC of 41.9%. around risk score of 80% b) Ventilator classification achieves a maximum MCC of 34.8%. c) Death within 30 days model obtains 25.4% MCC. d) ICU admission classifier reaches 38.1% MCC. Clinicians may select the optimal risk score thresholds based on their clinical need by understanding the relationship between sensitivity and sensitivity.

V. Discussion and Conclusion

In this work, we developed a risk score for four clinical outcomes using EHR data available at time of diagnosis to assist healthcare workers in prioritizing patients for monoclonal antibody treatment. The risk scores were developed using deep neural networks that achieved state-of-the-art performance for this task (Figure 5). We used a real-world clinical dataset with over 17,000 patients from the Emory University Hospital system. Available data for patient risk assessment include past diagnoses and prescriptions, demographics, and basic laboratory, and vital sign measurements.

Our initial training endpoint was the binary label, 'anyCatastrophic', which encompassed any patients that a) required mechanical ventilation, b) were admitted to the ICU, c) died within up to 90 days of their COVID-19 diagnosis, or d) were admitted to a hospital within 90 days of their COVID-19 diagnosis. Using this binary label, we trained a total of 9 different shallow learning models, of which Gaussian Naive Bayes had the best performance in the validation set, with an AUROC of 72.6%. Following this, we designed 11 different deep learning models with the different architectures shown in Figure 3, and selected the model with the best performance on the validation set (model 3, AUROC of 88.1%). This performance constitutes a 15.5% improvement over our best baseline classifier.

To determine the efficacy of our model to effectively detect future negative clinical outcomes, we separately assessed classification performance for the following: a) patients that required mechanical ventilation ('vent'), b) patients that died within 30 days after their COVID-19 diagnosis ('death30'), and c) patients that were admitted to the ICU ('ICU'). We examined model performance using AUROC, MCC, sensitivity, and specificity (Figures 4 and 5). Interestingly, for the test set, the AUROC of the more granular labels 'vent' and 'death30' (88.2% and 87.8%, respectively) were higher than the test set AUROC of 'anyCatastrophic' (86.7%), suggesting that the model is generalizable to labels that were not used during hyperparameter tuning.

Finally, an important decision was how to present the overall output of the classifier. While a binary output could simplify the output, we reasoned that it would be easier for clinicians to interpret if the output were to be provided as a numeric value from 0 to 1, obtained using the sigmoid activation function on the output layer of the neural network (Figure 2). This severity score can then be generated for each patient. Using a display, such as those shown in Figure 5, a clinician would be able to customize their use of the score to fit their goal (e.g. screening vs. medication administration).

There are several potential future directions to take this work. For example, it will be important to find ways to enhance the interpretability of the model. Out of the total of 56 features used by the model, it would be of interest to identify the most important ones used for each classification task. Methods such as local interpretable model-agnostic explanations (LIME) [9] or SHapley Additive exPlanations (SHAP) [10] can be used to establish feature importance.

Another potential expansion on the current work may be to make the model more robust to missing data. Several features had to be removed because they had a high percentage of missing data, but they could otherwise have been informative. Even for the features that were retained, imputation can create bias in the dataset by making assumptions on the underlying data distribution. One way to increase the robustness of the model might be to add dropout to the input layer during training, such that a random subset of the features are initially set to zero for each iteration. During testing of the trained model, if there are features with missing data, they can be set to zero without substantially affecting the performance, since the model would have already been trained to manage this issue.

Finally, we will seek to test the hypothesis that our approach can be generalized for use with other hospital datasets. This will involve obtaining de-identified EHR from COVID-positive patients from multiple healthcare systems. If successful, our models have the potential to help clinicians across institutions triage COVID-19 patients in order to provide more targeted care to those at highest risk for severe complications.

References

[1] N. Trivedi, A. Verma, and D. Kumar, "Possible treatment and strategies for covid-19: review and assessment," *European review for medical and pharmacological sciences*, vol. 24, no. 23, pp. 12 593–12 608, 2020.

[2] F. S. Heldt, M. P. Vizcaychipi, S. Peacock, M. Cinelli, L. McLachlan, F. Andreotti, S. Jovanović, R. Dürichen, N. Lipunova, R. A. Fletcher *et al.*, "Early risk assessment for covid-19 patients from emergency department data using machine learning," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[3] I. Landi, B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, and R. Miotto, "Deep representation learning of electronic health records to unlock patient stratification at scale," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[4] H. Estiri, Z. H. Strasser, J. G. Klann, P. Naseri, K. B. Wagholikar, and S. N. Murphy, "Predicting covid-19 mortality with electronic medical records," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–10, 2021.

[5] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli *et al.*, "Baseline characteristics and outcomes of 1591 patients infected with sars-cov-2 admitted to icus of the lombardy region, italy," *Jama*, vol. 323, no. 16, pp. 1574–1581, 2020.

[6] S. Richardson, J. S. Hirsch, M. Narasimhan, J. M. Crawford, T. McGinn, K. W. Davidson, D. P. Barnaby, L. B. Becker, J. D. Chelico, S. L. Cohen *et al.*, "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area," *Jama*, vol. 323, no. 20, pp. 2052–2059, 2020.

[7] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui *et al.*, "Clinical characteristics of coronavirus disease 2019 in china," *New England journal of medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[8] W. T. Li, J. Ma, N. Shende, G. Castaneda, J. Chakladar, J. C. Tsai, L. Apostol, C. O. Honda, J. Xu, L. M. Wong, T. Zhang, A. Lee, A. Gnanasekar, T. K. Honda, S. Z. Kuo, M. A. Yu, E. Y. Chang, M. ". R. Rajasekaran, and W. M. Ongkeko, "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Sep. 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-01266-z

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[10] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.