Using Machine Learning to Predict Hospitalization and Mortality of COVID-19 Patients with Diabetic Retinopathy

Katherine Zhong,[1,2] Elizabeth Chen, PhD,[1,2] Carsten Eickhoff, PhD,[1,2] Paul B Greenberg, MD, MPH[1,3]; on behalf of the N3C Consortium
[1]The Warren Alpert Medical School [2]Center for Biomedical Informatics, Brown University
[3]Section of Ophthalmology, Providence VA Medical Center

Purpose: The prognosis and epidemiology of severe COVID-19 illness in patients with diabetic retinopathy (DR) are not well understood. Using electronic health record (EHR) data from the National COVID Cohort Collaborative (N3C) Data Enclave, we performed a retrospective cohort study and tested the hypothesis that machine learning (ML) can be applied to a multi-center national dataset to build a predictive model that identifies risk factors for hospitalization and mortality of COVID-19 patients with DR.

Methods: We developed a random forest classifier model to identify patients at risk of hospitalization and mortality using EHR data from the N3C Data Enclave. We defined our base population (n= 31,419) as patients who have a DR diagnosis on or prior to their first positive COVID-19 lab result or diagnosis. Data were analyzed using computer programming languages including Python, PySpark, R, and SQL. 80% of the data were randomly assigned as the training set, and the remaining 20% were reserved as the test set. Random forest classifier models were built, and 100 features were identified to train the models, including demographic information, drug exposure, comorbidities, procedures, and lab measurements. Model performance was evaluated using area under the receiver operating characteristic curve (AUROC) on the test set. Feature importance was determined via Shapley values.

Results: The random forest classifier model achieved AUROC of 0.7082 for predicting hospitalization and 0.7940 for predicting mortality. Important risk factors for hospitalization included patient age, comorbidities (kidney disease, heart disease, chronic lung disease), and drug exposure. Important risk factors for mortality included lab measurements, patient age, and comorbidities. In addition, patients with DR and COVID-19 who present with a more advanced stage of DR and have other diabetic complications relative to those who have an early stage of DR and fewer diabetic complications were more likely to be hospitalized.

Conclusions: Our results suggest that ML can be applied to a large dataset to predict clinical outcomes. Our model reveals that age and lab measurements were the most important features in predicting COVID-related mortality, and the leading comorbidities of severe COVID illness in DR patients included kidney disease, heart disease, and chronic lung disease.

Funding resources: This work was supported by National Institutes of Health grants T35HL094308, U54GM115677, and U54GM104942.