Transparent and Distributed AI Prediction Modeling: A Case Study on Pediatric Covid

Fernando Suarez Saiz[1]*, Sanjoy Dey*[2]*, Prithwish Chakraborty*[2]*, Mohamed Ghalwash*[2]*, and Pablo Meyer* [2]
[1] *IBM Watson Health;* [2] *Center for Computational Health, IBM Research*

Background: COVID-19 pandemic has had heterogeneous impacts on pediatric populations with most of them presenting asymptotic or having mild symptoms, nevertheless some will progress to severe disease. One of the most severe conditions is the Multisystem Inflammatory Syndrome in Children (MIS-C), characterized by inflammation of multiple organs and tissues including the heart, lungs, blood vessels, kidneys, digestive system, brain, skin or eyes. Therefore, it is important for healthcare providers to determine such severe patient groups as early as possible so that appropriate interventions can be undertaken to improve pediatric patient outcomes. Machine learning models for predicting such outcomes early using Electronic Health Records (EHR) could potentially help clinical decision making such as advanced and more intensive medical interventions.

Methods: We aimed to develop a clinical risk assessment model for predicting a) COVID positive pediatric samples with higher risk of being hospitalized, and b) the risk of requiring ventilation or cardiovascular interventions. We used a recently proposed Pediatric Comorbidity Index (PCI) based on 27 predefined conditions and empirically identified conditions that provide a summary measure of disease burden in addition to the common features available in typical EHR datasets such as demographic data, 90-day diagnosis history and class of drugs. We used LightGBM [1] a highly efficient and low resource implementation of gradient boosting trees, to build our predictive models. Furthermore, for better patient level interpretability, we used a state-of-the-art model interpretation framework called SHAP [2] to identify the major risk factors associated with the complications. This combination thus allowed us to obtain both higher prediction power and model interpretation for both pediatric COVID-19 problems. Moreover, we conducted these analyses using Lightsaber [3], our recently open-sourced framework that automates data ingestion, model training, tracking, and evaluation aspects so that users focus on their model logic. Lightsaber allowed us to rapidly conduct several experiments over choices of hyper-parameter, conduct ablation studies to understand importance of features such as PCI, and ensure reproducible, distributed experiments with full audit trails to ensure transparency. In addition, the experiments were conducted in a closed computation enclave – using Lightsaber we were able to deploy our analysis developed elsewhere on the closed enclave without significant transfer effort.

Results: We used a large-scale pediatric COVID-19 cohort from the NCATS National COVID Cohort Collaborative (N3C) Data Enclave as a participant of the BARDA Community Challenge - Pediatric COVID-19 Data Challenge. The dataset consists of 6 million patients, including more than 2 million COVID positive patients, and nearly 7 billion rows of records. For the first task of predicting hospitalization, we applied Lightsaber framework to predict 2,162 hospitalizations out of 81,742 pediatric patients who tested COVID-19 positive in an outpatient setting. For the second task we aimed to predict the characteristics of 486 patients needed either ventilation or cardiovascular support out of 7,286 hospitalized patients. We tested the effectiveness of using PCI as a feature set in addition to other features like demographic, prescribed drugs and lab-tests. For the hospitalization prediction task, the model achieved area under the ROC curve (AUROC) of 0.84, area under the precision-recall curve (AUPRC) of 0.225, accuracy of 0.86 and F2 score of 0.33, whereas we achieved 0.991 AUROC, 0.925 AUPRC, accuracy of 0.95 and F2 score as 0.86 for the task of predicting ventilation and cardiovascular treatments. Important risk factors were identified by looking at the SHAP values of the obtained model. Results show that a higher PCI overall increases hospitalization risk as well as PCI_asthma and other commonly reported risk factors for COVID-19 plus geographic information.

For the second prediction task, the prominent risk factors contain PCI and higher representation of class of drugs described by their ATC codes such as Antineoplastic agents, Cardiac therapy that point to underlying conditions influencing the outcome and serve as confounders of an underlying illness but seemingly not child obesity. More importantly, the PCI was important to predict risk of ventilation use and specifically the conditions related to cardiovascular disease (as with the drugs) and anemia. We think that the implementation of this index is one of the major contributions to the explainability/predictions of our model.

Conclusion: Overall, we implemented a high-performing and interpretable model through Lightsaber for predicting pediatric COVID-19 severity. The risk factors derived from the PCI, the laboratory measurements in particular leukocyte counts, demographic data such as sex, age and weight are important to aid clinical decision making.

References:

1. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
2. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
3. Suryanarayanan, P., Chakraborty, P., Madan, P., Bore, K., Ogallo, W., Chandra, R., Ghalwash, M., Buleje, I., Remy, S., Mahatma, S. and Meyer, P., 2021. Disease Progression Modeling Workbench 360. *arXiv preprint arXiv:2106.13265*.