Table of Contents

[Timeline 3](#_Toc133484820)

[Challenge Administration 3](#_Toc133484821)

[*Participant Onboarding* 3](#_Toc133484822)

[*Challenge data management* 4](#_Toc133484823)

[*Team Resources* 4](#_Toc133484824)

[*Challenge Submissions* 4](#_Toc133484825)

[Task 1 Outcome Definitions 5](#_Toc133484826)

[Task 2 Outcome Definitions 6](#_Toc133484827)

[Summary of Submission Methods 7](#_Toc133484828)

[Model Evaluation 7](#_Toc133484829)

[**Evaluation Panel** 7](#_Toc133484830)

[**Qualitative Metrics** 7](#_Toc133484831)

[**Utility** 7](#_Toc133484832)

[Interpretability 7](#_Toc133484833)

[Timeliness 8](#_Toc133484834)

[Context Utility 9](#_Toc133484835)

[**Reproducibility** 12](#_Toc133484836)

[Technical Reproducibility 12](#_Toc133484837)

[Prediction Reproducibility 12](#_Toc133484838)

[Documentation 12](#_Toc133484839)

[Supplementary Analyses 12](#_Toc133484840)

[**Supplementary Analysis 1: MIS-C Sensitivity Analysis** 12](#_Toc133484841)

[**Supplementary Analysis 2: Measurement only evaluation for task 2** 13](#_Toc133484842)

[Winners and Honorable Mentions Methods 14](#_Toc133484843)

[Task 1 Winner - A LightGBM Model to Predict COVID-19 Pediatric Hospitalizations 14](#_Toc133484844)

[**Introduction** 14](#_Toc133484845)

[**Methods** 15](#_Toc133484846)

[**Feature Generation** 15](#_Toc133484847)

[**Model Development** 16](#_Toc133484848)

[**Discussion** 16](#_Toc133484849)

[**References** 17](#_Toc133484850)

[Task 2 Winners - Submission to the Pediatric Severity Prediction Challenge 17](#_Toc133484851)

[**Background/Introduction** 17](#_Toc133484852)

[**Methods** 18](#_Toc133484853)

[**Input data** 18](#_Toc133484854)

[**Feature engineering** 18](#_Toc133484855)

[**Classification** 19](#_Toc133484856)

[**Explainability** 19](#_Toc133484857)

[**Discussion** 19](#_Toc133484858)

[Honorable Mention – Clinical Utility 20](#_Toc133484859)

[**Background/Introduction** 20](#_Toc133484860)

[**Methods** 20](#_Toc133484861)

[**1) Feature Set** 20](#_Toc133484862)

[**2) Machine Learning Algorithm** 23](#_Toc133484863)

[**3) Feature Importance and Shapley-Value Analysis** 23](#_Toc133484864)

[**Conclusion/Discussion** 24](#_Toc133484865)

[**References** 26](#_Toc133484866)

[Honorable Mention – Feature Interpretability and Design 26](#_Toc133484867)

[An interpretable ensemble classifier for pediatric COVID-19 hospitalization can improve the utilization of monoclonal antibody therapy 26](#_Toc133484868)

[**Background/Introduction** 26](#_Toc133484869)

[**Methods** 27](#_Toc133484870)

[Feature Extraction 27](#_Toc133484871)

[Model derivation 27](#_Toc133484872)

[Model performance and interpretation 27](#_Toc133484873)

[**Conclusion/Discussion** 30](#_Toc133484874)

[**References** 31](#_Toc133484875)

[Honorable Mention – Computational Methodology 32](#_Toc133484876)

[Multi-model Ensemble AI can Accurately Predict Pediatric COVID-19 Significant Cardiac Intervention or Invasive Ventilation Need upon Hospitalization 32](#_Toc133484877)

[**Background/Introduction** 32](#_Toc133484878)

[**Methods** 32](#_Toc133484879)

[**Conclusion/Discussion** 35](#_Toc133484880)

[**References** 36](#_Toc133484881)

**Supplemental Materials**

# **Timeline**

The Pediatric COVID-19 Data Challenge began on August 19, 2021 with the Onboarding phase, a period of time to allow participants to form teams, gain access to the N3C enclave, and to begin experimenting with the available data. Round 1 of the challenge began on September 15, 2021, where participants began building models for Task 1 of the challenge. The training data for Task 1, along with the gold standard answers, was made available to all challenge participants. Round 2 of the challenge began on November 12, 2021, where the training data for Task 2 was made available. During this period, the training data for both Task 1 and Task 2 were made available simultaneously. Round 1 and 2 ended on December 17, 2021 when data and code access was removed from the teams. From December 17, 2021 to March 9, 2022, the Evaluation Phase was held, during which models were evaluated and scored. Task winners and honorable mentions were announced on April 5, 2022. (Supplemental Figure 1)



Supplemental Figure 1. Timeline of the Pediatric COVID-19 Data Challenge. The challenge was carried out from August 19, 2021 to March 9, 2022 in 3 phases: the onboarding phase, the challenge phase (rounds 1 and 2), and the evaluation phase.

# **Challenge Administration**

*Challenge.gov*

The Pediatric COVID-19 data challenge was executed through the challenge.gov website (https://www.challenge.gov/challenge/pediatric-covid19-data-challenge/) and was utilized to post information on the overview for the challenge, prizes, eligibility rules, judging criteria and onboarding instructions.

## *Participant Onboarding*

The administration and onboarding for the Pediatric COVID-19 Data Challenge was done through the synapse.org collaboration platform run by Sage Bionetworks. Registration, agreeing to the challenge rules and terms, team formation, distribution of rules and challenge resources, and the discussion forum were facilitated through the synapse landing page.

Separately, participants were onboarded into the N3C enclave by following the standard onboarding procedures laid out by the N3C organization, including having a signed Data Use Agreement between their host institution and N3C, receiving IRB approval if required by their home institution, agreeing to not download any data from the enclave without permission, and agreeing to the terms of use of the N3C enclave. Participants could then request access to the challenge project using the existing Data Use Request (DUR) associated with the challenge project. Once each user had fully registered through the Synapse landing page, the challenge administrator would approve the participants data use request. The user was then added to their team folder as a “project collaborator”, giving them permission to view, edit, and add materials to their team folder.

## *Challenge data management*

The model development and evaluation for the Pediatric COVID-19 Data Challenge was conducted using the National COVID Cohort Collaborative data enclave, a secure platform built using the Palantir foundry operating system. The N3C data enclave is organized into individual projects, where each project is associated with a specific data use request (DUR) and a specific level of data access (de-identified data or a limited dataset). Usually, a given project has access to the most up-to-date version of the available data in the N3C enclave at the specified level type. For this challenge, we created two projects: the organizer project and the challenge project. The organizer project was a standard N3C project that had access to all level 2 data in the N3C enclave. The challenge project was a custom project that did not have access to any of the N3C data. Instead, the challenge data was curated in the organizer’s project and annotated with a custom challenge permission tag. The curated data was moved into the challenge project and users associated with that project were allowed to access the custom challenge permission tag. This allowed the organizers to “freeze” the challenge data separate from the main enclave to limit the ability of the challenge participants to glean information about the testing dataset from the full N3C data. The challenge data was stored in the challenge project where all participants had access.

## *Team Resources*

Within the challenge project, each team was given a folder where they would build their models. The access configurations were setup so that only members of a team were able to edit and create files in their team folder. Every participant in the challenge could see other team folders and could read the file names, but they could not view or edit the contents.

## *Challenge Submissions*

Over the course of the challenge, teams were expected develop their prediction models using python, R, or SQL using the tools available in the N3C enclave platform. The final submissions were expected to take in two datasets, the training and testing dataset, which were each subsets of the full N3C data. Participants were expected to train their models on the provided training data and to make predictions on every patient available in the person table of the testing data (Supplemental Figure 2). Initially, the testing data was a placeholder dataset made up of 300 randomly selected patients from the training data. During the evaluation phase of the challenge, the testing dataset was replaced with the prospectively collected testing dataset.

Diagram

Description automatically generated

Supplemental Figure 2. Instructions for the format of challenge submissions. Team models were expected to intake two datasets, the training and testing datasets, and their code was expected to train a model and to use that model to make predictions about every patient in the person table of the testing dataset. The final output needed to be a single file with two columns: person\_id and a likelihood score.

Winner Announcement

After the conclusion of the challenge, BARDA announced the winners at <https://medicalcountermeasures.gov/stories/pediatricchallengedata/> and https://drive.hhs.gov/pedwinners.html.

# **Task 1 Outcome Definitions**

In task 1, participants were asked to address the following question: O*f pediatric patients who test positive for COVID-19 in an outpatient setting, who are at risk for hospitalization?*  The main goal behind this task was to have teams build predictive tools that could be used in an outpatient setting to assess the risk of a given pediatric patient progressing to a level of COVID-19 severity that warranted a trip to the hospital (including an emergency department encounter (Supplemental Table 4) or inpatient hospital visit (Supplemental Table 3). COVID-19 related outpatient visits were defined as outpatient visits (Supplemental Table 2) that occurred within 7 subsequent days of a patient’s earliest COVID-19 positive test. The presence of an inpatient visit (Supplemental Table 3) or emergency department visit (Supplemental Table 4) within 35 days of a COVID-19 outpatient visit was determined to be the true positive outcome.

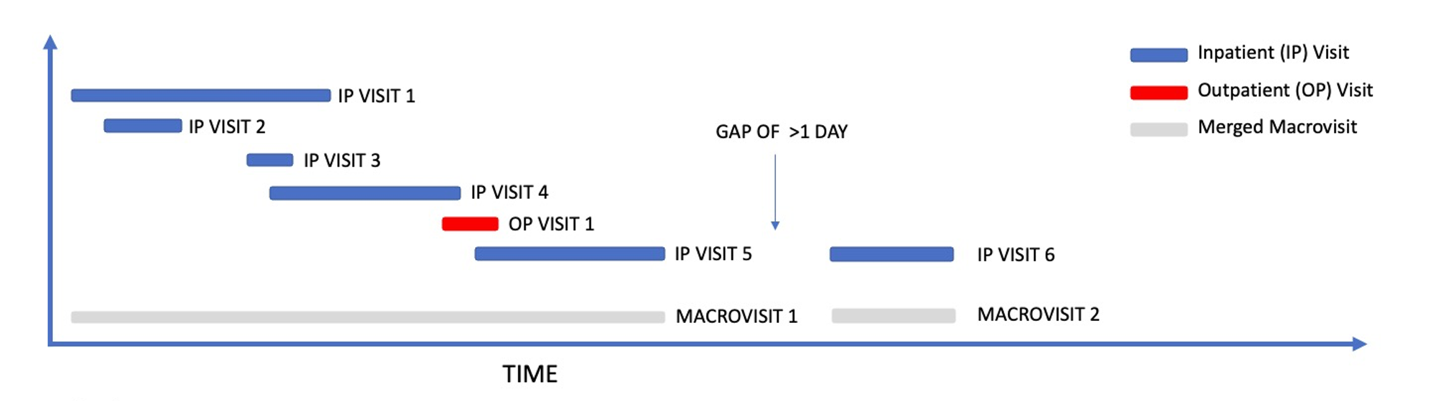
Patients with same day hospitalizations, where the hospitalization or emergency room visit occurred on the same day as the outpatient visit were excluded. This was due to the difficulty in differentiating between clinical information collected during the outpatient visit versus the inpatient or emergency department visit. Models that were developed in this task needed to be able to make their predictions using only clinical information available at the time of the outpatient visit. While datetime is available, its reliability is questionable, clinical records can be recorded hours after clinical observation, and often, no time is recorded. Thus, the cleanest solution was to exclude same day hospitalization and emergency room visits.

# **Task 2 Outcome Definitions**

In task 2, COVID-19 positive hospitalized patients were defined as any patients who tested positive for COVID-19 within 7 days of being admitted to an inpatient visit. The true positives included all patients who, during their hospitalization, needed mechanical ventilation (Supplemental Table 5), Extracorporeal Membrane Oxygenation (ECMO) (Supplemental Table 6), cardiovascular support (Supplemental Tables 7-14), or passed away during their hospital stay.

Cardiovascular support was defined as any patient who was given the injectable forms of Dobutamine (Supplemental Table 7), Dopamine (Supplemental Table 8), Epinephrine (Supplemental Table 9), Levosimendan (Supplemental Table 10), Milrinone (Supplemental Table 11), Norepinephrine (Supplemental Table 12), Phenylephrine (Supplemental Table 13), or Vasopressin (Supplemental Table 14). The death table available in the OMOP Common Data Model was used to identify which patients passed away.

To define a hospital stay, macrovisits were used to identify overlapping visits in each patient record to indicate the patient was under continual clinical care. A macrovisit was defined as a merger of all overlapping or sequential inpatient patient visits, to which any other types of visits (OP, telephone, etc) were added that occurred during the merged interval. (Supplemental Figure 3)



Supplemental Figure 3. Visual example of how macrovisits are defined. Multiple visits that have overlapping dates are combined to define the full macrovisit, where the start and end date of the macrovisit are the earliest and latest dates of the overlapping visits. All other clinical record information can be mapped to a macrovisit.

# **Summary of Submission Methods**

The vast majority of teams used XGBoost as their method, with regression, naive bayes, nearest neighborhood, and neural network methods represented as well. Teams used feature engineering techniques such as curated concept sets available in the N3C enclave, clinical expertise to guide data curation, and a “grab all the features” approach. Top models mostly used concept sets or ontology roll ups to create their features and many had clinicians review the concepts for accuracy and relevance to the prediction tasks. Boosting methods are computationally cheaper than many other methods, which may have driven their usage. While the N3C enclave is a powerful platform, it is still limited in resources and developing larger, more computationally expensive models can be challenging.

# **Model Evaluation**

## **Evaluation Panel**

The evaluation panel was made up of subject matter experts from medical diagnostics, clinical, statistical, and informatics domains. A subset of the evaluation panel were the judges who were responsible for assigning scores to the submitted models in accordance with a qualitative evaluation rubric. Each judge scored the models across six qualitative metrics.

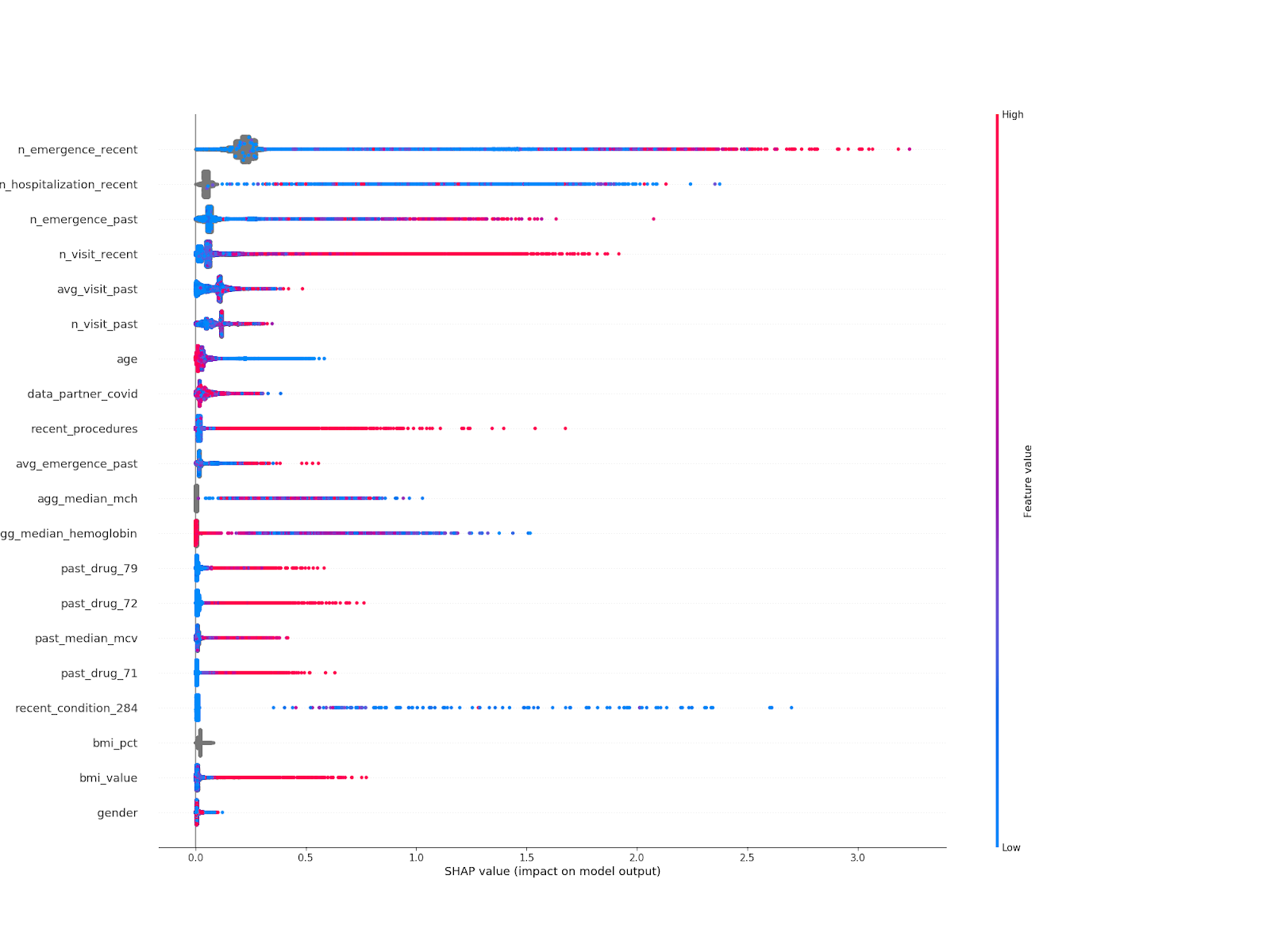
## **Qualitative Metrics**

The qualitative metrics were broken down into two broad categories, **utility** and **reproducibility**. Under utility, subject matter experts evaluated the models for interpretability, timeliness, and context utility. Under reproducibility, subject matter experts evaluated models for technical reproducibility, prediction reproducibility, and documentation. For each metric, the judges scored the models on a scale from 0-5.

### **Utility**

### *Interpretability*

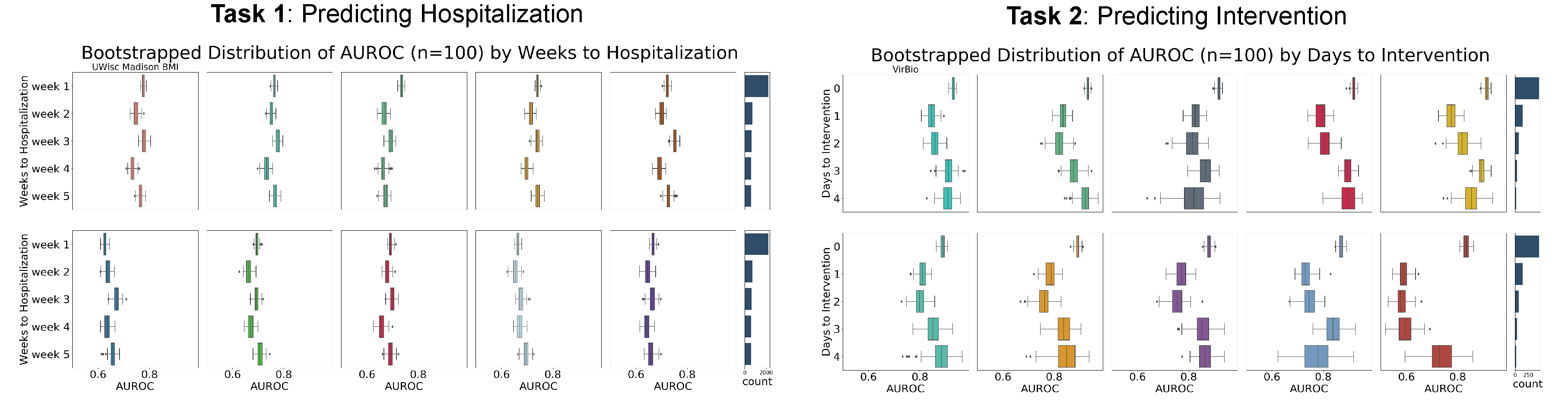
When looking at interpretability, the judges considered how easily understandable the submitted models were, taking into account the model calibration curves, reported features and the associated weights, clarity of the reported methods, and visualizations created by the teams to aid in interpretation of the factors driving model performance (Supplemental Figure 4). Significant time was spent looking at reported features and their importance. Aside from understanding what features were being used, evaluators considered whether each feature made sense for the given prediction problem.



Supplemental Figure 4. Example visualization of feature importance from UWisc-Madison-BMI's submission from task 1. The visualization was created by UWisc-Madison-BMI as part of their final submission write up and represents the SHAP values on each of their most impactful features. Blue indicates a low feature value, and red represents a high feature value. The further to the right the point, the higher the impact that feature had on the individual model prediction.

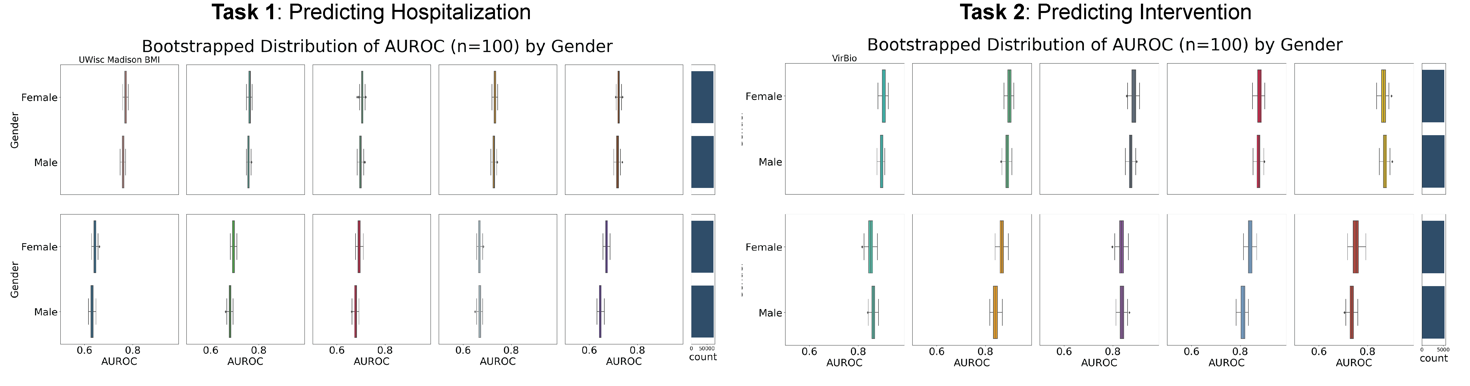
### *Timeliness*

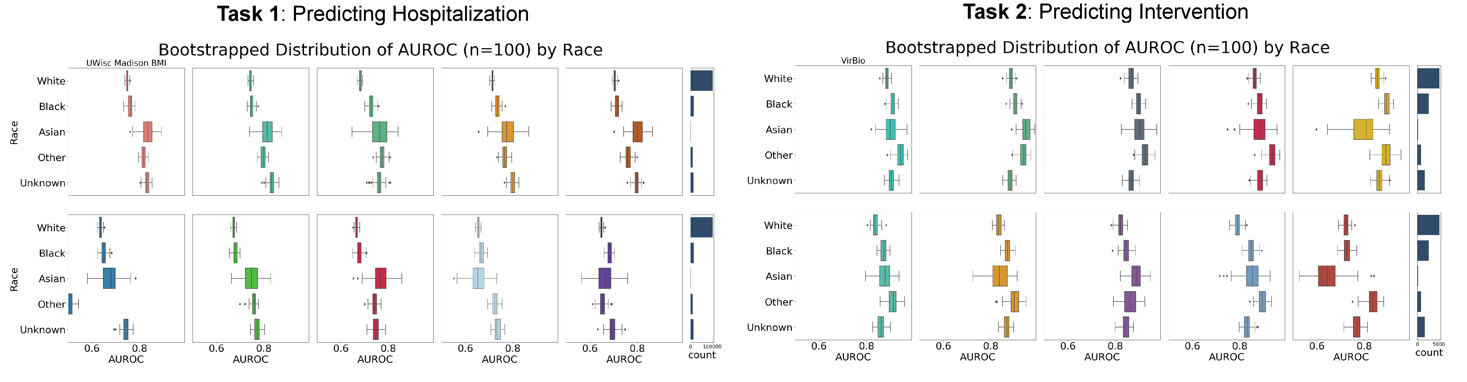
When reviewing timeliness, the evaluators considered how far out models were able to accurately predict the outcomes. For task 1, evaluators looked at weekly times bins and for task 2, evaluators looked at daily time bins. Supplemental figure 5 shows the bootstrapped distributions of the top 10 models for each of the tasks across the time windows of interest. In task 1, evaluators looked to see how far out models were accurately predicting hospitalization from the time of the outpatient visit. In task 2, evaluators looked to see how far out models were accurately predicting mechanical ventilation or cardiovascular support from the time of hospitalization. Similar patterns emerged, with task 1 showing a distinct “W” pattern that was repeated across most models, indicating an underlying pattern of hospitalization that was data driven.

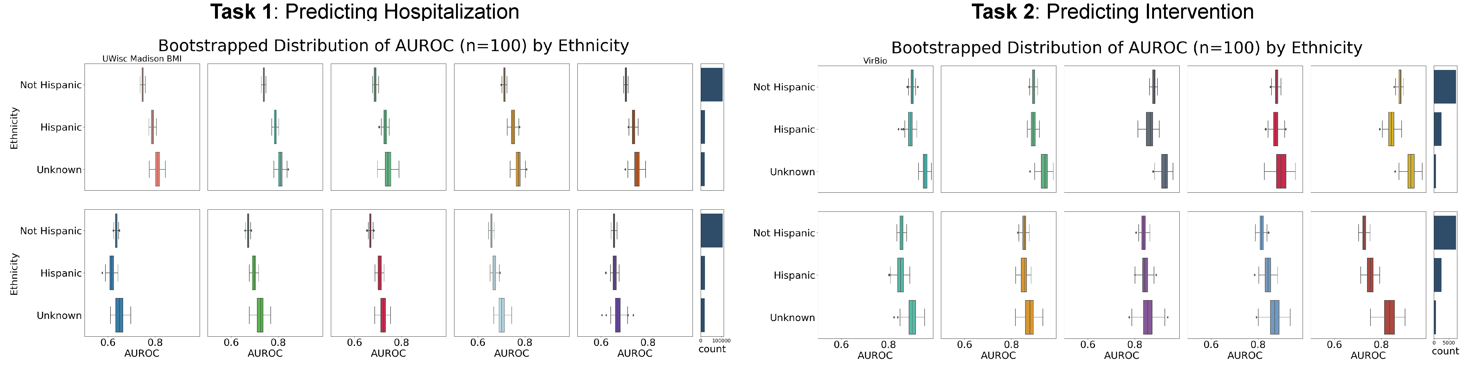
Supplemental Figure 5. Time to outcome assessments for the top 10 quantitative scoring models in task 1 and task 2. All team names have been removed except for the winning teams. Outcome time bins are defined as non-cumulative counts of patients who have the outcome in question within the given time window. Patients included in week one or day zero are not included in week four or day four and are excluded from AUROC calculation. The boxplot represents the bootstrapped distribution of the AUROC (n=100) for the given time window. The bar chart on the right shows the number of patients who have the outcome of interested during the given time window.

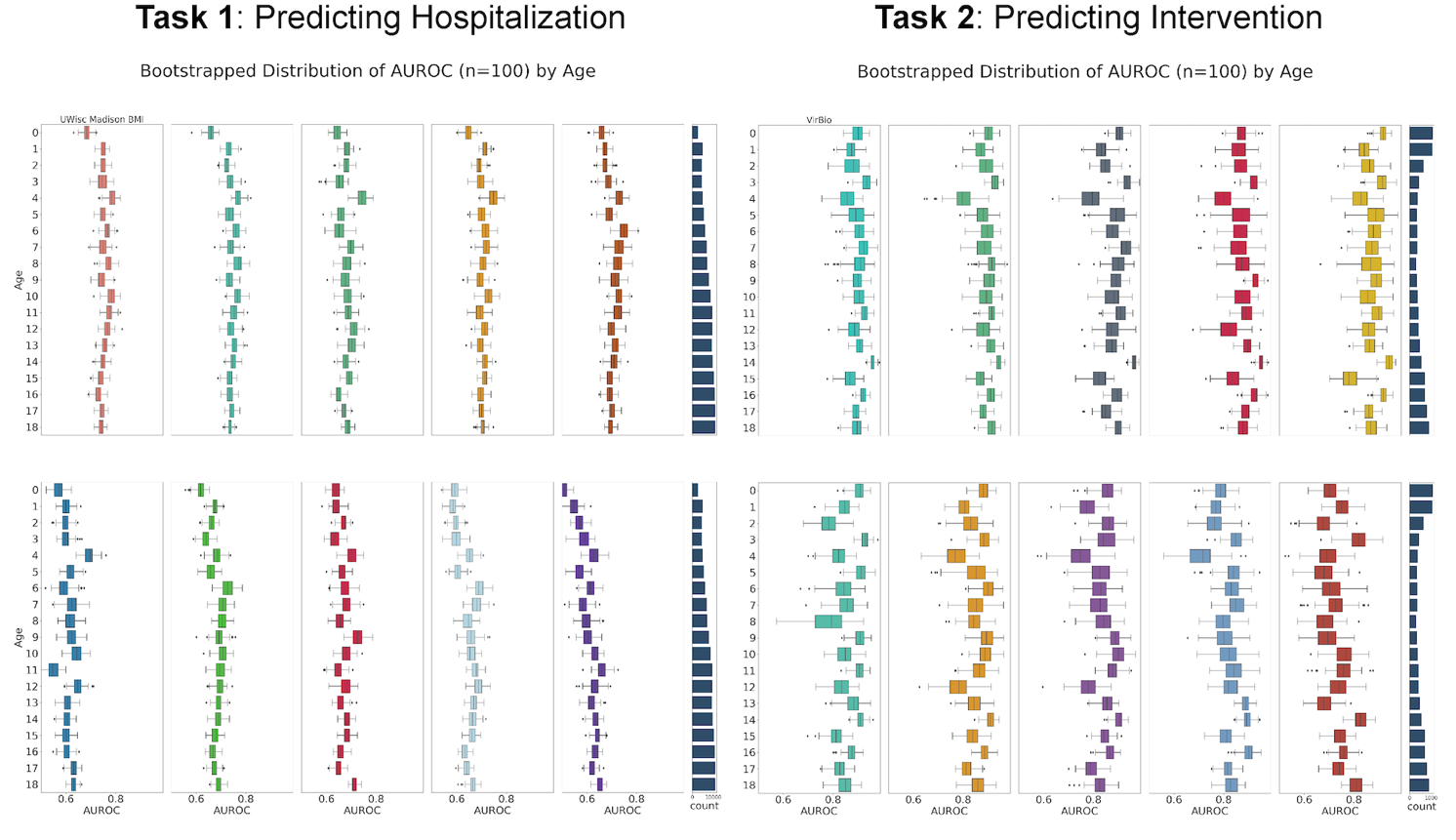
### *Context Utility*

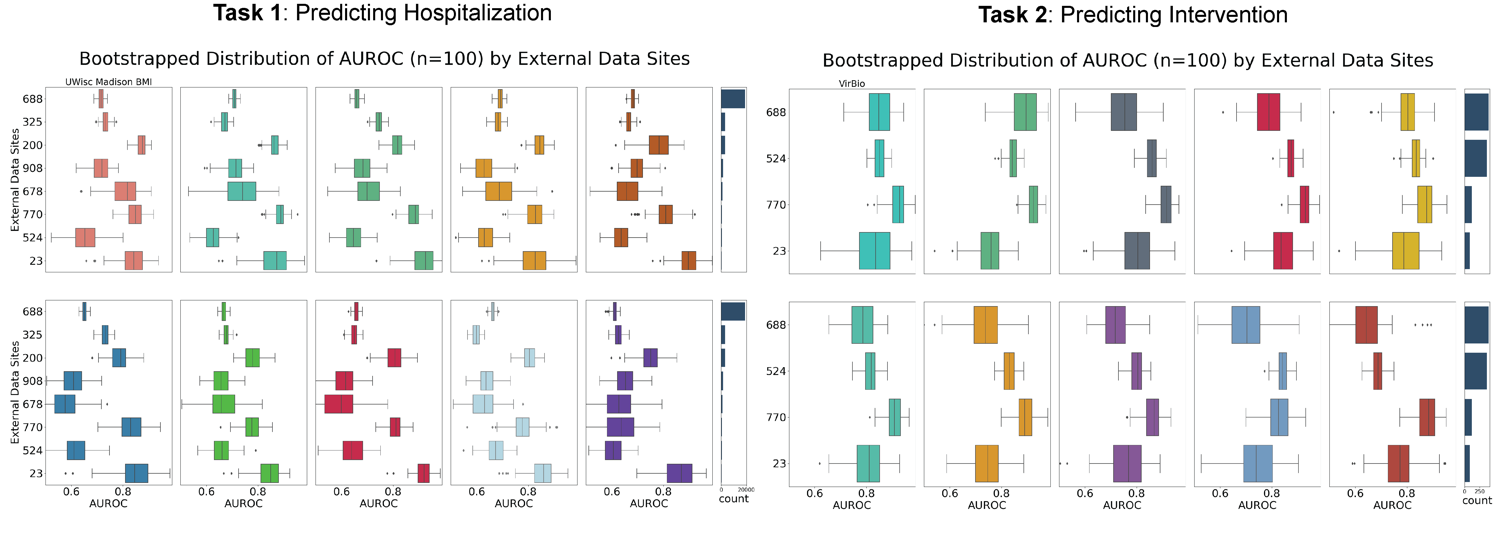
When reviewing models for context utility, the evaluators looked at model performance under a variety of circumstances, including looking at how models performed across different demographic strata, clinical phenotypes, and data from newly onboarded data sites in the N3C enclave that had been added during the course of the challenge. Supplemental figures 6-9 show the bootstrapped AUROC distributions (n=100) of the top 10 highest scoring models for each of the tasks for different demographic strata (Gender, Race, Ethnicity, Age). Supplemental figure 10 shows the bootstrapped AUROC distributions (n=100) of the top 10 highest scoring models for each task on each newly onboarded data contributing site. These evaluations were used by each judge to derive a score of context utility.

Supplemental Figure 6. Bootstrapped distribution of the AUROC (n=100) on the evaluation cohort broken down by gender for the top 10 quantitative scoring models in task 1 and task 2 across gender. Non male or female genders made up a tiny fraction of all patients and so were excluded from this analysis. All team names have been removed except for the winning teams. The bar chart on the right shows the number of patients who are labeled as that gender.

Supplemental Figure 7. Bootstrapped distribution of the AUROC (n=100) on the evaluation cohort broken down by race for the top 10 quantitative scoring models in task 1 and task 2 across gender. All team names have been removed except for the winning teams. The bar chart on the right shows the number of patients who are labeled as that race.

Supplemental Figure 8. Bootstrapped distribution of the AUROC (n=100) on the evaluation cohort broken down by ethnicity for the top 10 quantitative scoring models in task 1 and task 2 across ethnicity. All team names have been removed except for the winning teams. The bar chart on the right shows the number of patients who are labeled as that ethnicity.

Supplemental Figure 9. Bootstrapped distribution of the AUROC (n=100) on the evaluation cohort broken down by age for the top 10 quantitative scoring models in task 1 and task 2 across gender. All team names have been removed except for the winning teams. The bar chart on the right shows the number of patients who are labeled as that age.

Supplemental Figure 10. Bootstrapped distribution of the AUROC (n=100) on the evaluation cohort broken down by source data center for the top 10 quantitative scoring models in task 1 and task 2. All team names have been removed except for the winning teams. The bar chart on the right shows the number of patients who are labeled as that gender. External data sites are data sites that were onboarded into the N3C enclave over the course of the challenge and so are “external” to the challenge training data. Sites that are included in the task 1 evaluation but not included in the task 2 evaluation did not have a sufficient number of severe COVID-19pediatric patients to be included in task 2.

### **Reproducibility**

### *Technical Reproducibility*

The challenge administrator evaluated each model for how robust they were to data changes.. Common causes of errors included incorrectly accounting for new or missing variables in the held out test set, out of memory errors when models did not account for the size of the new testing data, and incorrect data input formats. Judges evaluated the number of changes that needed to be made to the code to run.

### *Prediction Reproducibility*

Models were rerun on a subsample of the testing data to see if model training workflows were deterministic, or at a minimum, barely shifted in their prediction scores. The subsample included 300 test set patients that were held constant for all evaluated models. The training data remained the same. Models were re-trained and re-run against the 300 subsampled test set and individual patient-level model prediction scores were compared to the original prediction scores given to those patients. Judges evaluated deviations from the original prediction scores and the variance of those differences..

### *Documentation*

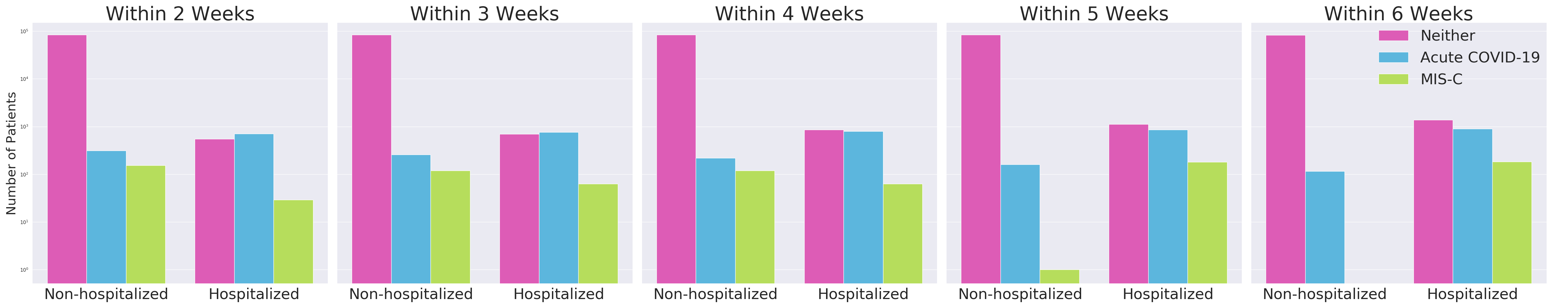
Evaluators were asked to review the commented submitted code and final submission write up for clarity and understandability. High scoring code submissions for this section included block of code with descriptive comments, both in the code itself as well as the team write up that was submitted along with the code. High scoring final submission write ups clearly outlined the methodology used, including data curation, feature engineering, and model selection, as well as including rationale for using chosen methods.

# **Supplementary Analyses**

## **Supplementary Analysis 1: MIS-C Sensitivity Analysis**

In task 1, one of the goals from submitted models was to be able to predict which patients were at risk for hospitalization from Multisystem Inflammatory Syndrome in children. While a diagnosis code exists in ICD10 to identify MIS-C patients (M35.81), the phenotype capture in the EHR is very low. In order to diagnose MIS-C, every other alternative possible diagnosis needs to be ruled out, which means a confirmed diagnosis does not always appear in the records. In addition, often times the MIC-S diagnosis is located in the clinical notes and is not part of the tabular data. Electronic phenotypes exist to identify MIS-C patients, but these have the potential to introduce artifacts when you use them in benchmarking exercises. The concern was that participants would end up predicting or recreating the electronic phenotype, rather than building a clinically relevant model on the confirmed true positive outcomes.

The desire was to have task 1 models predict far enough into the future that some of the predicted hospitalizations were MIS-C patients and acute COVID-19 patients, but not so far that the non-COVID-19 related hospitalizations made up the vast majority of hospitalizations. The existing M35.81 ICD10 diagnosis code as well as the Acute COVID-19 concept sets (Supplemental Table 16) were used to identify COVID-19 related hospitalizations. The assessment examined how far out from the initial COVID-19 adjacent outpatient visits the COVID-19 related hospitalizations were still occurring. After six weeks, MIS-C related hospitalizations disappeared, but acute COVID-19 hospitalizations still appeared, while within the 5-week window, MIS-C related hospitalizations were still occurring (Supplemental Figure 11). Based on this information, we defined task 1 as risk of hospitalization within five weeks of the covid adjacent outpatient visit.

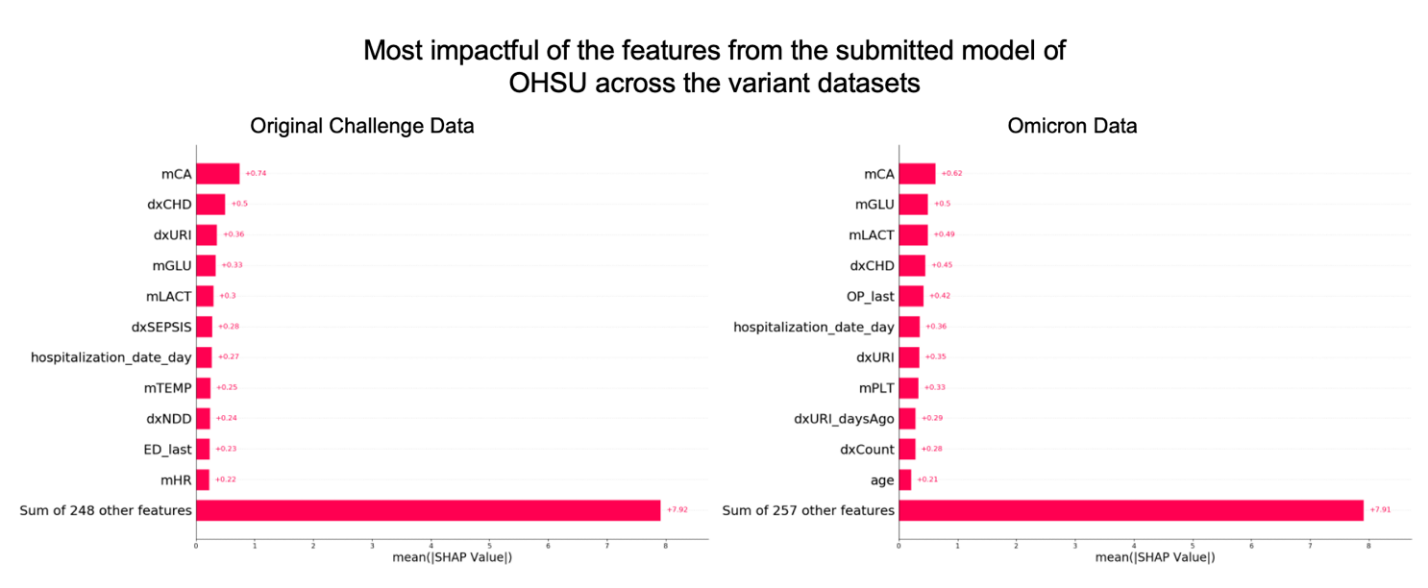
Supplemental Figure 11. Analysis of all COVID-19 positive pediatric patients who tested positive for COVID-19 within seven days of an outpatient visit and the number of patients who are hospitalized with either a MIS-C diagnosis code or an acute COVID-19 diagnosis code within different time windows. Each subplot is comparing the number of patients who have no hospitalization and have a MIS-C, acute COVID-19, or neither diagnosis code to the number of patients who have been hospitalized within the designated time period and have a MIS-C, acute COVID-19, or neither diagnosis code. The Y axis is in log scale. Each subplot is looking at a given time window defined from the outpatient visit date of interest.

## **Supplementary Analysis 2: Measurement only evaluation for task 2**

When evaluating task 2, the clinical concepts that were used to derive the gold standard outcomes were removed from the record on the day of a patient’s COVID-19 related hospitalization; however, tangential, or semi-related concepts were not removed. During initial evaluation of the models, many models were simply using semi-proxy codes as indicators for the core gold standard outcomes. To evaluate the robustness of models in a clinical setting, all records from the day of initial hospitalization and after the day of initial hospitalization, except for measurements coded to that date were removed for evaluation. All models decreased in performance, however, the top model, VirBio, remained the top performing model and had one of the smallest drops in performance.

**Supplementary Analysis 3: Omicron variant data evaluation post-challenge**

The models from the two winners and the honorable mentions were run on a prospectively collected dataset from the time when the Omicron variant was the most prevalent strain of COVID-19. Models were trained on the original challenge training dataset and applied to the newly collected Omicron-era dataset. The dataset was evaluated for UWisc-Madison-BMI, Oregon Health & Science University, and the Wind city team for Task 1 and Vir Bio, Oregon Health & Science (OHSU), and ARI Science team submissions for Task 2 (Figure 1). For Task 1, none of the models saw a significant decrease in their performance, and in the case of UWisc-Madison-BMI and Wind City, their models saw a non-significant increase in their performance (0.016 and 0.006 respectively). For task 2, the Oregon Health & Science University saw a significant drop in performance (-0.167), while Vir Biotech and ARI Science saw a small increase in their performance (0.012 and 0.014 respectively). To evaluate OHSU’s large drop off between the two datasets, OHSU’s feature impact was compared between the challenge data with Delta as the dominant strain and the post-challenge data with omicron as the dominant variant. Their model was trained and fit to the omicron-dominant data and their most impactful features were evaluated. Many features remained similar in their impact, namely serum calcium(mCA), diagnosis of congenital heart disease (dxCHD), serum glucose (mGLU), and serum lactate (mLACT) while new features appeared more impactful on the omicron data such as age, blood platelet count (mPLT), and last outpatient visit (Supplemental Figure 12).

Supplemental Figure 12. Comparison of the most impactful features of the OHSU model across the Original challenge data and the Omicron data. The original challenge data includes the original training and test data used to train and evaluate models during the challenge. The omicron data was collected after the challenge from the same data contributing sites, but only included patients who were added between December 9, 2021 and April 7, 2022.

# **Winners and Honorable Mentions Methods**

# **Task 1 Winner - A LightGBM Model to Predict COVID-19 Pediatric Hospitalizations**

Jifan Gao1, 2, Guanhua Chen1, 2

1 Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, USA

2 Contact: J.G. ([jifan.gao@wisc.edu](mailto:jifan.gao@wisc.edu)), G.C. ([gchen25@wisc.edu](mailto:gchen25@wisc.edu))

## **Introduction**

We adapt our winning solution of the first EHR DREAM Challenge - Patient Mortality Prediction [1, 2] to predict COVID-19 pediatric hospitalization. These two challenges are similar in the following aspects: they both use the OMOP data model, and they both are about predicting the onset of a certain clinical risk within a specific future time window. For these reasons, we use a similar LightGBM model to make probabilistic predictions and utilize the longitudinal data by ontology-rollup and time binning. Furthermore, there is also data shift caused by heterogeneities among different care sites in addition to that caused by time. In addition to ontology-up used in [2], we extract the medians of percentiles of synonymous concepts to represent measurement values and use concept sets built by N3C communities to reduce dimensionality. To ensure the generalizability of the model across different sites, we split the training data by *data partner ID*sto tune and validate the LightGBM model.

## **Methods**

### **Feature Generation**

We map condition concept IDs to ICD-10 subgroups and drug concept IDs to ATC pharmacological subgroups. The mapping tables can be constructed using [NIH's SNOMED-ICD10 mapping files](https://www.nlm.nih.gov/healthit/snomedct/us_edition.html) and the OMOP model's *concept\_ancester* table. This ontology-rollup strategy creates 288 mega concepts for medical conditions and 95 mega concepts for drugs. We count the occurrence of ICD-10 mega concepts for each patient within three days ahead of the covid\_index and before three days ahead of outpatient\_visit\_start\_date separately, which leads to 576 (288\*2) features. We count the occurrence of ATC mega concepts for each patient within three days ahead of the outpatient\_visit\_start\_date and before three days ahead of the outpatient\_visit\_start\_date separately, which leads to 190 (95\*2) features.

For numeric measurements, we first rank percentiles of each selected concept and take the average of synonymous concepts as mega concept representations. For example, when a patient has the measurements of Hematocrit [Volume Fraction] of Blood by Impedance (concept\_id: 3013752), and Hematocrit [Volume Fraction] of Blood by Automated count (concept\_id: 3023314) and the corresponding percentiles are 0.65 and 0.75, we use 0.70 to represent the hematocrit level of this patient. We select 18 mega measurements that are related to COVID severity [3] and construct three features for each measurement, including the most recent values of the mega concept values before the outpatient\_visit\_start\_date, the median of the mega concept values (i.e., the median of the medians of a group of synonymous concepts over time) before patients' outpatient\_visit\_start\_date, and the contrast between the recent values and the past median values. Such a strategy leads to 54 (18\*3) features.

We count the occurrence of COVID symptoms posted on the [CDC website](https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html) within three days before the covid\_index. The counts of occurrence of each complex chronic condition [4] before the covid\_index are also collected. We extract the history of diabetes, obesity, history of COVID, and influenza vaccination. We also record two types of procedures encountered prior to the outpatient\_visit\_start\_date: those encountered within 30 days before the outpatient\_visit\_start\_date, and those encountered 30 days or earlier separately. In addition, we collect demographic variables (such as age, race, and gender), total hospital visits, and the average number of visits per year for each patient. These strategies contribute to another 45 features.

In summary, we construct 865 features for our prediction model building.

### **Model Development**

We order data partner IDs by the number of first COVID diagnoses included from lowest to highest. Data from the first 20% data partners with the fewest COVID diagnoses are used as the testing set. The next 16% data partners with the fewest COVID diagnoses are used to tune the model, and the remaining data are used to train the model. The purpose is to mimic the situation in which there are a variety of unseen data sources in the evaluation stage. We use Optuna [5] for automatic parameter tuning and a LightGBM [6] model for prediction. After finding optimal hyperparameters using the above approach, we train a new model with all the training data with such hyperparameters and use this model to make predictions in the evaluation stage. We use the SHAP [7] package to obtain and visualize the feature contribution in our model. The evaluation metric is AUPR + optimal F2 + cross-site AUROC as defined in the Challenge Instruction. We do not take the square of each term to mitigate the scale difference of the three individual scores.

## **Discussion**

When we evaluated the model on our selected testing set, the AUPR, optimal F2, and the cross-site AUROC are 0.155, 0.302, and 0.753 respectively. There are a few directions that may improve the model's numerical performance. First, timestamps (except the birth of year) in this dataset are randomly shifted within 180 days due to the consideration of data security and privacy, the interpretability of the model may be harmed if we collect the (shifted) date of the COVID diagnosis even if we find these shifted dates can slightly improve the numerical performance. Second, we do not use ensemble for better interpretability in our work, although it is known that the ensemble trick may also increase the model's numerical performance. Third, we expect incorporating COVID vaccination information can improve the model's predictive power; however, that is not the case for this task as less than 0.1% of patients have at least one COVID vaccination record before their first COVID diagnosis. We suspect that COVID vaccination information is not fully captured in the dataset. Our model's performance can be further improved if the complete COVID vaccination information is available in the dataset.

On our testing set, below are the top 10 features that make the most contribution to our model's predictions:

* the number of emergence visits within the past year before the outpatient\_visit\_start\_date.
* the number of hospitalization in the past year before the outpatient\_visit\_start\_date.
* the number of emergence visits within the past before the outpatient\_visit\_start\_date.
* the number of total hospital visits within the past year before the outpatient\_visit\_start\_date
* the average of hospital visits per year.
* the number of total hospital visits within the past year before outpatient\_visit\_start\_date.
* Age
* ID of the data partner in which the COVID diagnosis is reported.
* the total number of procedures within three days before the outpatient\_visit\_start\_date.
* the average of emergence visits per year.

The 11th to 20th include measurements such as BMI, mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), historical drug exposure such as ATC subgroup "Drugs for obstructive airway diseases" (denoted as drug\_79), "Psychoanaleptics" (denoted as drug\_72), " Psycholeptics drugs" (denoted as drug\_71), and medical conditions such as ICD-10 subgroup "Persons with potential health hazards related to family and personal history and certain conditions influencing health status" (denoted as condition\_284) within three days before covid\_index. The SHAP values indicate that patients' general health status is the dominating factor for the risk prediction, and geographical factors (indicated by data partner ID), lab measurement, medical conditions, and drug history also play a role.

## **References**

* EHR DREAM Challenge - Patient Mortality Prediction (syn18405991).
* A LightGBM Model to Predict 180 Days Mortality Using EHR Data (syn21598356).
* Ballaz SJ, Pulgar-Sánchez M, Chamorro K, Fernández-Moreira E, Ramírez H, Mora FX, Fors M. Common laboratory tests as indicators of COVID-19 severity on admission at high altitude: a single-center retrospective study in Quito (ECUADOR). Clinical Chemistry and Laboratory Medicine (CCLM). 2021 Mar 5.
* Feudtner C, Feinstein JA, Zhong W, Hall M, Dai D. Pediatric complex chronic conditions classification system version 2: updated for ICD-10 and complex medical technology dependence and transplantation. BMC pediatrics. 2014 Dec;14(1):1-7.
* Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. InProceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining 2019 Jul 25 (pp. 2623-2631).
* Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30:3146-54.
* Mokhtari KE, Higdon BP, Başar A. Interpreting financial time series with SHAP values. InProceedings of the 29th Annual International Conference on Computer Science and Software Engineering 2019 Nov 4 (pp. 166-172).

# **Task 2 Winners - Submission to the Pediatric Severity Prediction Challenge**

Istvan Bartha, Cyrus Maher, Amalio Telenti

Vir Biotechnology Inc.

## **Background/Introduction**

We approach the problem as a machine learning task on tabular data, i.e. we strive to build a feature matrix of N rows and M columns where N is the number of samples and M is the number of columnar features. The EHR data represents three challenges to conventional tabular data analysis: a) it is a time series b) it is extremely sparse c) it is extremely diverse in the sense that similar concepts have different encodings.

To deal with the time axis of the underlying EHR data we bin the timestamped EHR records into a small number of time bins relative to the cutoff date (e.g. average blood pressure of a patient 4 to 30 days before the cutoff). To deal with the sparsity of the resulting feature matrix we employ a missingness aware gradient boosted tree classifier which learns both data and missingness patterns simultaneously, rendering data imputation unnecessary. To deal with the diversity of the representation of the concepts we use both manual curation and count-based integration of data belonging to related concepts (e.g. instead of taking the average blood pressure, we count the number of blood pressure records in a time period).

## **Methods**

### **Input data**

We use the EHR records from the following tables: location, visit\_occurrence, drug\_era, payer\_plan\_period, procedure\_occurrence, observation, condition\_era, measurement.

### **Feature engineering**

#### **Derived features**

* The calculated age, race and ethnicity attributes found in the person table are used as is.
* The location values are used as is.
* Each payer\_concept\_id found in the payer\_plan\_period table is used as binary indicator variable.
* The total number of records in the observation table is used as a feature.
* The total number of records in the measurements table is used as a feature.
* The total number of records in the condition\_era table is used as a feature.
* The total number of records in the drugs table is used as a feature.
* The total number of records in the visit\_occurrences table is used as a feature.
* The total number of records in any table after the cutoff (hospitalization/outpatient visit) date is predicted from the data available before the cutoff data. This predicted record volume is used a feature.
* We compute derived features like the ratio of neutrophils over lymphocytes from records in the measurement table and use this ratio as a feature
* We compute the time between the covid index and the date of outpatient visit or hospitalization.
* Based on the *concept\_set\_members* table we manually curate and harmonize the 83 conditions, observations and measurements, see Table 1 for the list.
* We also use as a feature the total number of records related to each concept\_set where the given concept\_set has at least 3 concept members.
* We time bin all the above listed features where applicable (if the record is time stamped).
* We also use the number of days between the cutoff date and very first record for each of the above listed features.
* We also use the number of days between the cutoff date and last record before the cutoff for each of the above listed features .

All features having less than 20 (for the features derived from the counts of concept\_sets) or five (the rest) non-missing values are dropped.

The above listed features total to more than 20 thousand. We employ a single variate feature selection loop to select a small number of them. We use features in the final multivariate setting which pass a single variable ROC AUC > 54%.

The final feature matrix for Task 1 has 1744 columns.

The final feature matrix for Task 2 has 2509 columns.

We evaluate the trained Task 1 model on the Task 2 instances and use that output as an additional feature for Task 2.

#### **Time binning**

All timestamped records are binned into the following time bins:

* the event happened zero and eight days before the outpatient visit
* the event happened between eight and 32 days before the outpatient visit
* the event happened 32+ days before the outpatient visit

For Task 2 we use 0-4, 4-8, and 8+ intervals in the past from the hospitalization date.

### **Classification**

We use the sklearn.ensemble.HistGradientBoostingClassifier for all training and prediction needs. This classifier is treating the missingness patterns as a feature, therefore we use no imputation.

### **Explainability**

To gauge internal performance we use the cross\_validate method in sklearn.

To gauge internal feature importances we use the permutation based features importance technique from the sklearn package on ROC-AUC values as metric. We used this to guide our feature engineering efforts and to drop certain record types from the Task 2 training.

We link the Task 1 and Task 2 feature importance scores into this document (Table 2. and 3.). Features whose permutation based standard deviation was too high (more than twice the mean) are not included in the table. While interpreting this table we warn for certain caution since none of the explainability methods we know of are free from issues of misinterpretation (e.g. <https://arxiv.org/pdf/2002.11097.pdf>, <https://arxiv.org/pdf/2002.11097.pdf>) .

## **Discussion**

For Task 2, we found that there are records originating from the first day of hospitalization which are very telling about a potential severe disease already present on that first day (certain procedures or certain drugs administered). In the current submission we replicated the task 2 ground truth script and remove all records occurring on the day of hospitalization and related to those concept ids which define the ground truth. However the feature importance of the submitted model reveal that there is still considerable information leakage in the retained training data: top influential variables are related to concepts of organ failure and critical care medications.

# **Honorable Mention – Clinical Utility**

Bruce L. Cragin, PhD1

1 Wind City Applied Research

## **Background/Introduction**

The N3C Enclave’s existing code sets and Codeset Explorer provide powerful tools for organizing data that can be distributed among tens, hundreds, or even thousands of different concept\_ids into a single, coherent, more useful whole. Generating feature sets that take advantage of this capability was a priority goal of this effort. The need to accommodate future training data updates with as yet not fully understood properties (resulting from new variants, increased utilization of vaccines, possible evolution from pandemic to endemic status, etc.) makes many-featured models an attractive choice. However, the use of many features also raises the likelihood of feature collinearity, which can diminish the performance and interpretability of logistic regression and similar models. Tree-based models are known to be highly immune to collinearity effects, and modern-day gradient boosted tree models have been among the top performers in predictive modeling competitions. The xgboost or "extreme boosting" model, in particular, handles missing values automatically using a method called sparsity-aware imputation that often provides improved performance relative to traditional imputation techniques. The current version of xgboost also incorporates code from the treeshap Shaply-value algorithm, providing a powerful technique for feature performance evaluation. The WCAR Task 1 and Task 2 models are very similar, both using R language xgboost models, and both employing the same 42 mostly codeset-indexed features.

## **Methods**

### **1) Feature Set**

The feature set used in these models comprises demographic data and various codeset-derived features including drug exposures, conditions diagnosed, laboratory test results, vital sign data, various OMOP table record counts, as well as four custom/special purpose features.

The demographic features are **pseudo\_age**, **MALE, FEMALE, Hispanic\_or\_Latino**, **Not\_Hispanic\_or\_Latino**, **Asian, Black\_or\_African\_American**, and **White**. The data needed to construct these features are contained in the appropriate goldstandard and person input tables. The pseudo\_age feature is defined as the time in months from the midpoint of the patient's birth year (nominally July 2nd) to the Covid index date. The ethnicity, gender, and race-related features are defined in a simple but straightforward way as 1-hot features. Patients whose entry for the ethicity\_concept\_name is neither "Hispanic" nor "Non-Hispanic" are coded as "Unknown", as are those whose gender concept name was neither "MALE" nor "FEMALE". Patients having race\_concept\_name "Black" are coded as 1 for "Black or African American" along with those directly matching that name. Unless the patient's race\_concept\_name is either "Black", "Black or African American", "Asian", or "White", the race-related 1-hots are coded as "Unknown". As usual, no separate Unknown 1-hot features are required.

The non-demographic features utilize data from the drug\_exposure, condition\_occurrence, measurement, device\_exposure, observation, or visit\_occurrence tables. Active time "spans" are defined for each table type, and all records with timestamps lying *outside* the range {(outpatient\_visit\_start\_date - span) up to and including the outpatient\_visit\_start\_date} for Task 1 [or, for Task 2, {(hospitalization\_date - span) up to and including the hospitalization\_date}] are *dropped* prior to feature extraction. For both Task1 and Task 2, span = 360 days is used for the drug\_exposure, condition\_occurrence, and observation tables, and span = 180 days is used for the measurement, device\_exposure, and visit\_occurrence tables.

Drawing on the remarkable collaborative power of the N3C Enclave codesets, 13 binary and 13 continuous features are defined. The binary features utilize data from either the drug\_exposure or the condition\_occurrence OMOP table. The drug\_exposure related binary features and their associated codesets are:

**NSAID** ("NSAIDs", Richard Zhu, codeset id = 640043362),

**CORT** ("ARIScience - Drug Corticosteroids Systemic", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 571514014),

**sys\_antiviral** ("Systemic Antiviral Medications", Davera Gabriel, codeset id = 245992624),

**EmCARD** ("Emergency Cardiac Medications", Tianyi Zhang, codeset id = 277534971), and

**insln\_mtf** (the set union of {[DM]Insulin, Carolyn Bramante, codeset id = 91074072} and {[DM]Metformin, Victor Garcia, codeset id = 677036102}).

Each of the above features is assigned the value 1 if the patient has at least one record remaining in the drug\_exposure table with a drug\_concept\_id contained in the indicated code set and the value 0 otherwise.

The 8 condition\_occurrence related binary features and their associated codesets are:

**asthma** ("asthma\_nmh", Daniel Meza, codeset id = 879212923),

**T1DM** ("Type 1 diabetes mellitus", Anna O'Malley, codeset id = 3420519),

**sepsis** ("[covid19 v1]Sepsis" , Benjamin Amor, codeset id = 795540721),

**resp\_dis** ("ARIScience-Respiratory Disorder", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 961458756),

**card\_dis** ("ARIScience-Cardiac", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 535122844),

**lung\_dis** ("ARIScience-Lung Disorder", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 720920942),

**electrolyte\_dis** ("ARIScience-Electrolyte Disorder", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 441231383) and

**metabolic\_dis** ("ARIScience-Metabolic Disorder", [joy@ariscience.com](mailto:joy@ariscience.com), codeset id = 741619680).

Each of these features is assigned the value 1 if the patient has at least one record remaining in the condition\_occurrence table with a condition\_concept\_id contained in the specified code set and the value 0 otherwise. The remaining 13 continuous (or quasi-continuous) codeset-derived features are of four general types that can be classified as "laboratory test results", "vital sign", "table record count" and "custom/special purpose".

The lab test result features and the codesets from which they are derived are:

**max\_AST** ("Aspartate aminotransferase (AST component)", Janos Hajagos, codeset\_id = 248480621),

**max\_BUN** ("Blood urea nitrogen (component)", Janos Hajagos, codeset\_id = 139231433),

**max\_CRP** ("CRP C reactive protein", Richard Moffitt, codeset\_id = 371622342, except concept\_id 4264024),

**max\_D\_dimer** ("D-dimer(components)", Benjamin Amor, codeset\_id = 475972797),

**max\_eGFR** ("eGFR", Jung Su, codeset\_id = 929126024),

**min\_pH** ("N3C pH of Blood ATLAS #927", Andrew Girvin, codeset\_id = 845428728), and

**min\_RBC** ("ARIScience-RBC Nucleated", [joy@ariscience.com](mailto:joy@ariscience.com), codeset\_id = 380300160).

The 5 vital sign features and the codesets from which they are derived are:

**max\_height** ("Height (LG34373-7 + SNOMED)", [tjla@novonordisk.com](mailto:tjla@novonordisk.com), codeset\_id = 186671483),

**max\_weight** ("Body weight (LG34372-9 and SNOMED)", [tjla@novonordisk.com](mailto:tjla@novonordisk.com), codeset\_id = 854721978),

**max\_sys\_bp** ("Systolic BP (LG33053-6 + SNOMED)", Janos Hajagos , codeset\_id = 186465804),

**min\_O2sat** ("Arterial oxygen saturation v2 Atlas# 879", Kate Bradwell, codeset\_id = 780678652), and

**max\_resp\_r** ("Respiratory Rate: RRv3 atlas#872(v2)", Tell Bennett, codeset\_id = 286601963).

The maximization or minimization for the lab test and vital sign features is performed over all matching records in the measurement table. Data from the "harmonized\_value\_as\_number" column are used whenever available; otherwise, the "value\_as\_number" column is used. Data for the lab test and vital sign features are particularly sparse. Missing values are represented as NA and imputation is performed later (internally) by the xgboost algorithm. Body mass index was not used as a feature here, in part because the sparseness of this data limits its usefulness and in part because the use of BMI in a pediatric context is nontrivial (see eg. Sandler, 2021). We obtained approval to incorporate WHO and CDC conversion tables in our models to allow the transformation of weight and height to appropriately age-adjusted Z-scores, but that substitution did not improve the cross-validation score and was therefore not retained in the final version.

The 5 table record count features are defined as follows:

**device\_count**: the number of records for the patient in the device\_exposure table in that table's time span.

**drug\_count**: the number of records for the patient in the drug\_exposure table in that table's time span.

**meas\_count**: the number of records for the patient in the measurement table in that table's time span.

**obs\_count**: number of records for the patient in the observation table in that table's time span.

**visit\_count**: number of records for the patient in the visit\_occurrence table in that table's time span.

Finally, the 4 custom/special purpose features are defined as follows:

**partner\_training\_mean**: For each unique data\_partner\_id in the training set, the mean value of the outcomes is calculated. This defines the partner\_training\_mean feature for the training data set. If the data\_partner\_id of a test set patient is one that was also present in the training set, the test feature is assigned that same mean value. Otherwise, the test patient feature is crudely imputed by assigning it a value equal to the median of all the known training-set means.

**n\_prior\_neg**: This feature estimates the number of prior negative Covid-19 test results each patient has received by counting the number of measurement table records for which the measurement\_concept\_id = 706163 and the value\_as\_concept\_name is "Negative" or "Not detected". The n\_prior\_neg value for a patient may be an indicator of their perceived risk of exposure.

**n\_vaxed**: This feature estimates the number of times the patient has been vaccinated against Covid-19 by counting the number of measurement table records that contain a measurement\_concept\_id within the ("DG COVID 19 Vaccine", Davera Gabriel, codeset\_id = 705273705) codeset. Oddly, only a tiny fraction (less than 0.1%) of the patients in the current training set have a nonzero value for this feature, but one may reasonably expect this to change with future updates.

**autumn\_winter**: Following the observation of Nichols et al. (2021) that at temperate, northern latitudes, seasonal coronaviruses tend to occur more frequently in November through March than in other months, a binary feature "autumn\_winter" is defined with value 1 if the month of the outpatient\_visit\_start date [or in the case of Task 2, the hospitalization\_date] lies in this range, and zero otherwise. For the current training dataset, this feature has little if any predictive power, but if Covid-19 were to evolve from pandemic to endemic status it could eventually prove useful as a predictor of seasonal recurrence.

### **2) Machine Learning Algorithm**

The R language xgboost algorithm is well documented in Chen and Guestrin (2016). Model hyperparameters used here for tuning were the iteration step size eta**,** tree dimension max\_depth, sub-sampling fraction subsample, detuning/regulator factor gamma**,** and (maximum allowed) tree size nrounds. Seat-of-the-pants optimization was employed using the precision-recall and ROC area-under-curve scores obtained from 5-fold, early-stopping-enabled xgb.cv cross-validation calculations as a guide. Values thus selected for Task 1 were eta = 0.05, subsample = 0.5, nrounds = 400, max\_depth = 3 and gamma = 0. [The values used for Task 2 were exactly the same except that max\_depth = 2 was selected.] In cross-validation, this model yielded a ROC-AUC score of approximately 0.80 and a PR-AUC score of approximately 0.14. [The corresponding results for the much easier Task 2 were 0.90and 0.62]. Further tuning of the hyperparameters would improve these results slightly at the expense of increased execution time, but substantial improvement is not expected.

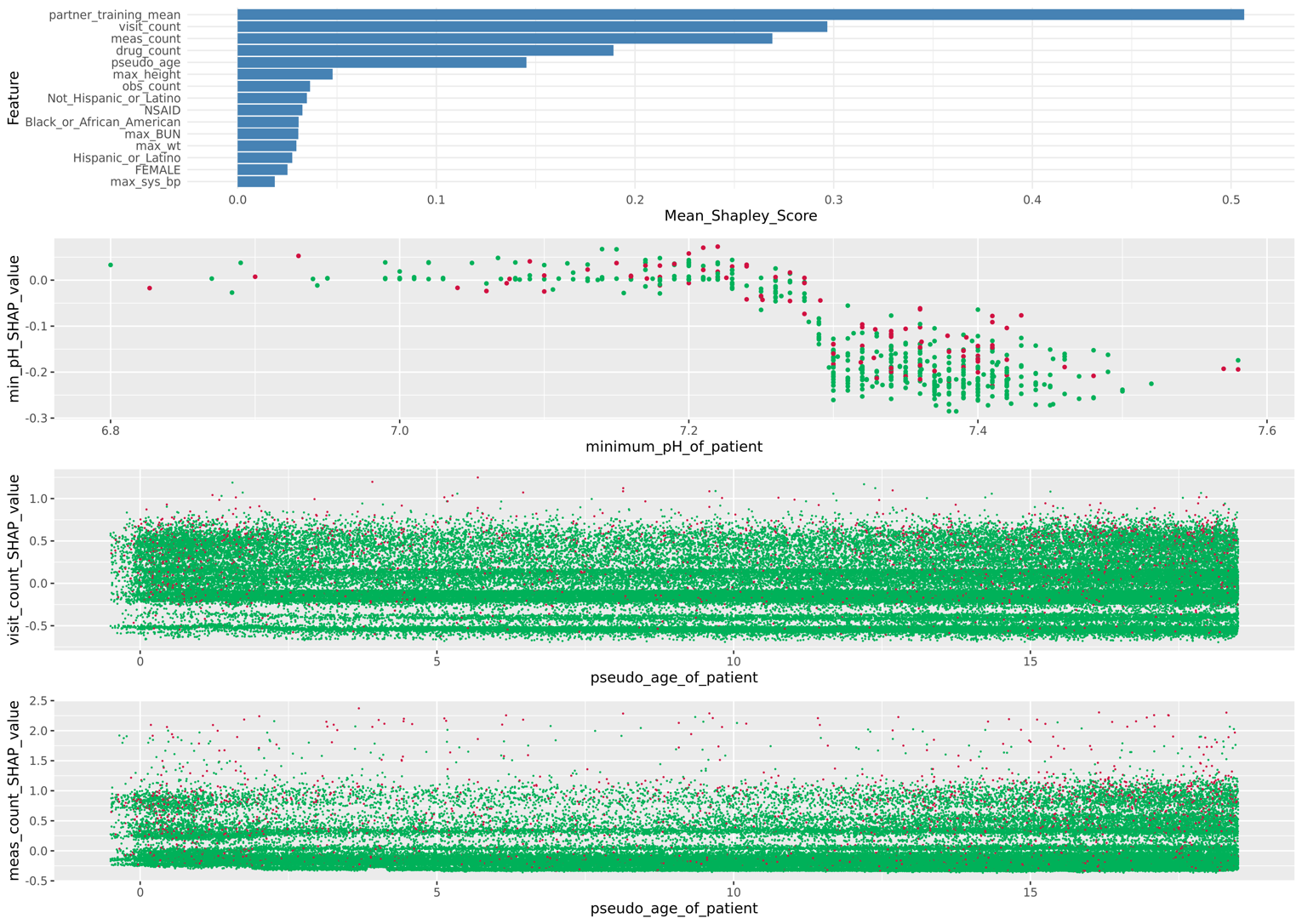
### **3) Feature Importance and Shapley-Value Analysis**

SHAP(SHapley Additive exPlanation) analysis (e.g. Merrick and Taly, 2020) can be thought of as generalizing the concept of a vector of the feature importances of a model to a feature importance matrix each row of which is applicable to an individual patient. SHAP matrix coefficients can be either positive or negative according to whether the associated feature contributes more to the separation of T from F positives or to the separation of T from F negatives. By averaging the absolute values of the matrix over patients by feature, one obtains so-called Mean SHAP Scores. These provide highly repeatable and mathematically well-justified measures of feature importance. Xgboost provides the SHAP value matrix via a simple call to its predict method, facilitating the use of this powerful method. Time does not permit a lengthy discussion here of the SHAP results for the models, but it is of interest to look at a few examples. This will be done in the following section.

## **Conclusion/Discussion**

The top panel of the figure below shows the mean SHAP scores obtained for the top 15 features for Task 1. The single most important feature was the institution-level one, "partner\_training\_mean". The predictive power of this feature probably derives either from regional differences or from a range of different types of institutions that contribute data to the enclave. The second, third, and fourth most important features were those derived from record counts over the visit\_occurrence, measurement, and drug\_exposure tables. One expects these features to be at least weakly correlated with each other and to serve as crude measures of the patient's general state of health. The patient's pseudo\_age was the fifth most important feature, with a mean SHAP score of approximately 0.15. This feature accounts for about 10% of the total predictive power of the top 5 predictors. The next 10 features, including max\_height, obs\_count, NSAID use, max\_BUN, max\_weight, max\_sys\_bp, and various demographic factors, had relatively low SHAP values. A complete list of the feature importances can be found near the end of the Logs tab of the train\_infer tab in the code workbook.

**Train infer**

The next three panels in the figure show examples of "SHAP scatter plots" that demonstrate the way in which the importance of individual features varies with the characteristics of the patients considered. The red dots here represent patients with positive outcomes and the green dots negative. The first example is for the maximum blood pH feature, max\_pH, which for Task 1 was the 28'th most important feature overall. For Task 1 this feature has predominantly negative SHAP values, extending from -0.25 for patients with normal or above normal pH and rising to near zero for patients with pH values well below normal. This shows that *normal* pH values are indicative of an increased likelihood of a *negative* outcome. Of note, the corresponding plot for Task 2 has a similar, sigmoidal shape, but in that case, the curve is shifted upward to almost entirely *positive* values. This shows that for Task 2, *low* pH is a strong indicator of a *positive* outcome, while normal and above normal pH values have less predictive power.

Panels 3 and 4 of the figure show how two of the most important Task 1 features, vis\_count and meas\_count, vary in their predictive power as a function of patient age. The variation is rather subtle, but there is at least some indication that both features become more powerful for patients less than about 2 or over about 12 years of age.

## **References**

* Sandler, Austin, The legacy of a standard of normality in child nutrition research, SSM - Population Health, Volume 15, 2021, 100865, ISSN 2352-8273. <https://doi.org/10.1016/j.ssmph.2021.100865>.
* Nichols, G.L., Gillingham, E.L., Macintyre, H.L. et al. Coronavirus seasonality, respiratory infections and weather. BMC Infect Dis 21, 1101 (2021). <https://doi.org/10.1186/s12879-021-06785-2>
* Chen, Tianqi and Guestrin, Carlos, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust 2016 Pages 785–794. <https://doi.org/10.1145/2939672.2939785>
* Merrick L. and Taly A. (2020) The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science, vol 12279. Springer. <https://doi.org/10.1007/978-3-030-57321-8_2>

# **Honorable Mention – Feature Interpretability and Design**

# **An interpretable ensemble classifier for pediatric COVID-19 hospitalization can improve the utilization of monoclonal antibody therapy**

Meenakshi Mishra (1), Benjamin Orwoll MD MS (1,2), Lorne Walker MD PhD (1,3)

1. Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Sciences University
2. Division of Pediatric Critical Care Medicine, Oregon Health and Sciences University
3. Division of Pediatric Infectious Diseases, Oregon Health and Sciences University

## **Background/Introduction**

The COVID-19 pandemic has been a devastating source of morbidity and mortality since its identification in December 2019 [1]. Children infected with SARS-CoV2 have low rates of hospitalization and severe disease relative to their adult and elderly counterparts [2]. In the setting of an ongoing pandemic, this presents a challenge to the healthcare system. How can we best provision therapies (such as monoclonal antibodies), vaccinations, healthcare resources and research funding to best serve children at the highest risk for severe COVID-19? Task 1 of the Pediatric COVID-19 Data Challenge addresses this via development of models predicting hospital admission after SARS-CoV2 infection.

This classification task presents several challenges. First, there is significant class imbalance - very few children are admitted to the hospital after COVID-19 infection. This is exacerbated by the observation that COVID-19 is often well-tolerated, even amongst higher-risk groups (children with cancer, diabetes, transplant recipients) [3]. Model precision may be limited as many ostensibly high-risk children do not require hospitalization. Second, admission data is not specific to COVID-19. Children may be admitted to a hospital after SARS-CoV2 infection for non-related reasons (a broken leg, for instance), creating heterogony amongst model targets. Finally, complex "black box" models may not provide sufficient context for reliable and unbiased use in a clinical setting. In our submission for Task 1, we seek to overcome these challenges with accurate and interpretable predictions of hospitalization risk after SARS-CoV2 infection.

## **Methods**

### *Feature Extraction*

Our model is based on data included in the person, condition\_era, measurements, visit\_occurrence, and drug\_era data tables. We identified concept sets represented in the concept\_set\_table to consolidate disparate data representations into parsimonious variables with clear clinical interpretations. All concepts that map to a known or likely risk factor for hospitalization (e.g. Type 1 diabetes, albuterol use, vital sign values) were mapped to a consolidated concept set to enhance interpretability [4]. Existing concept sets were reviewed by one of the challenge team members for suitability, and new concept sets were created using the concept set editor as necessary (refer to Concept\_sets\_and\_Ids for a full data dictionary). Measurement data was filtered to data in the 180 days prior to the index outpatient visit and corrected to SI units as needed. Additional features representing lag time to most recent diagnosis, measurement and drug records were derived from data tables. Missing data was assigned to the population median, or a pre-designated value (eg. fraction of inspired oxygen at 21%) as appropriate. After binary encoding of categorical variables, 262 features were candidates for model inclusion. For post-hoc analysis, hospitalizations were additionally labeled as COVID-related or non-COVID based on the presence or absence of a COVID-19 diagnosis code associated with the hospital encounter.

### *Model derivation*

Several standard machine learning models were considered for use: regularized logistic regression (LR), random forests (RF), neural networks (NN), boosted decision trees (XGB). Exploratory model development was performed by splitting the supplied training data 80%/20% into a development and validation set. Model selection and hyperparameter tuning were performed via 5-fold cross-validation on the development set to maximize area under the receiver-operating characteristic (ROC) curve and evaluated against the validation set. We noted that each of these models misclassified overlapping but different sets of individuals in the validation data set. We therefore evaluated an ensemble classifier comprised of a weighted mean of probability estimates from each candidate model. Weights for the ensemble classifier were also tuned as hyperparameters. After model development, the full training set was used to perform hyperparameter fine-tuning with 5-fold cross-validation and to fit final versions of the LR, RF NN and XGB models. The final model was an ensemble of LR, RF, and XGB models with weights 2, 4 and 3 respectively. The empirically determined weight for the NN model was 0, and it was excluded from the final ensemble classifier.

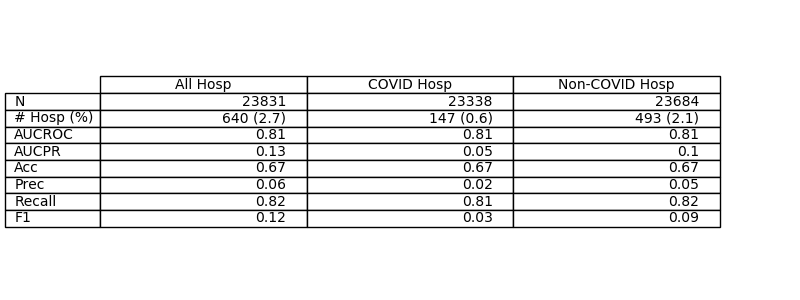
### *Model performance and interpretation*

Model performance was estimated by the performance of the development models on the validation set. Area under the ROC curve and precision-recall (PR) curve were 0.81 and 0.13 (Figure 1) respectively. Additional model statistics are summarized in table 1 (cutoffs for accuracy, precision, recall and F1 score given by the threshold with the maximum Youden index = sensitivity + specificity -1).

We used SHapley Additive exPlanations (SHAP) to provide model-wide as well as individualized characterizations of the features explaining model predictions [5]. The TreeExplainer Shapley method was applied to the XGB model, and both population-average and individual model explanations were derived. Additionally, we performed *post hoc* stratification of the validation set to evaluate model characteristics on specific sub-groups. Overall model performance and average feature importance were similar on sub-groups determined by admission reason, (COVID vs non-COVID), gender, race and ethnicity. The main exception to this observation was an increased importance of membership in Data Partner #399 for children with Hispanic ethnicity (Figure 2). A closer examination reveals that this data partner is comprised of 7,699 children, of which 59% are Hispanic (versus 19% in the full data set) and is associated with zero hospitalizations. The complete lack of hospitalizations likely represent a data collection artifact, but may have important implications for the performance of this model on the Hispanic population.

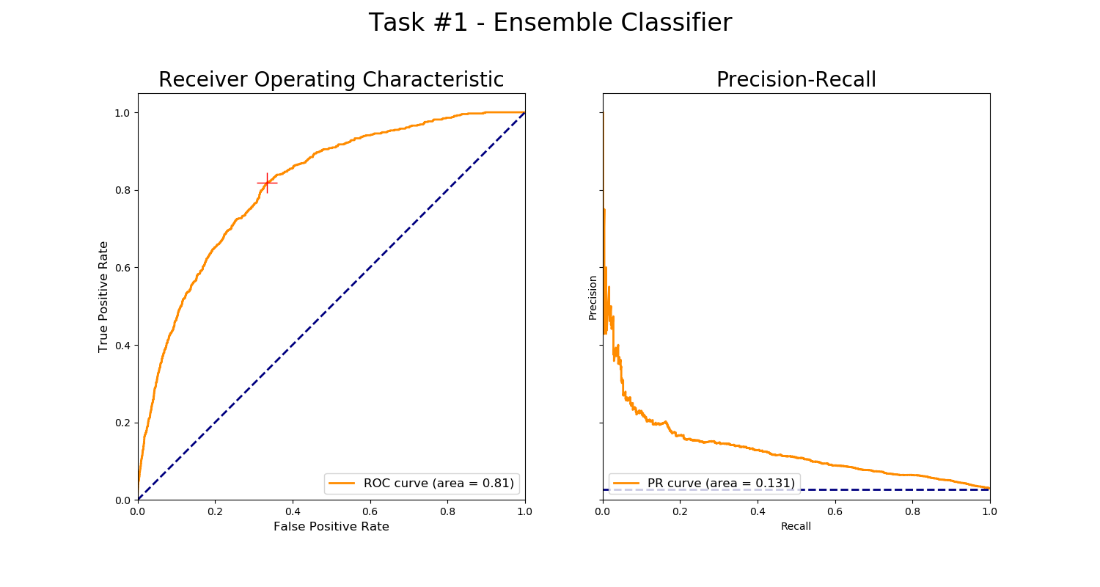
Finally, to characterize one potential use case for this model we explored the hypothetical impact of using model predictions to prioritize children for outpatient treatment with monoclonal antibody (mAb) therapy. In adult literature, this therapy has been shown to reduce risk of hospitalization when given in the outpatient setting [5]. Under the assumption of a 2.7% overall admission rate, and a 2/3 reduction of hospitalization risk with mAb therapy, we explored the likely impact of targeting this therapy to high-risk children. mAb therapy is a costly treatment in terms of healthcare utilization, requiring a prolonged infusion encounter with risk of anaphylaxis, therefore optimizing healthcare delivery is clinically important.

**Table 1: Model performance statistics**

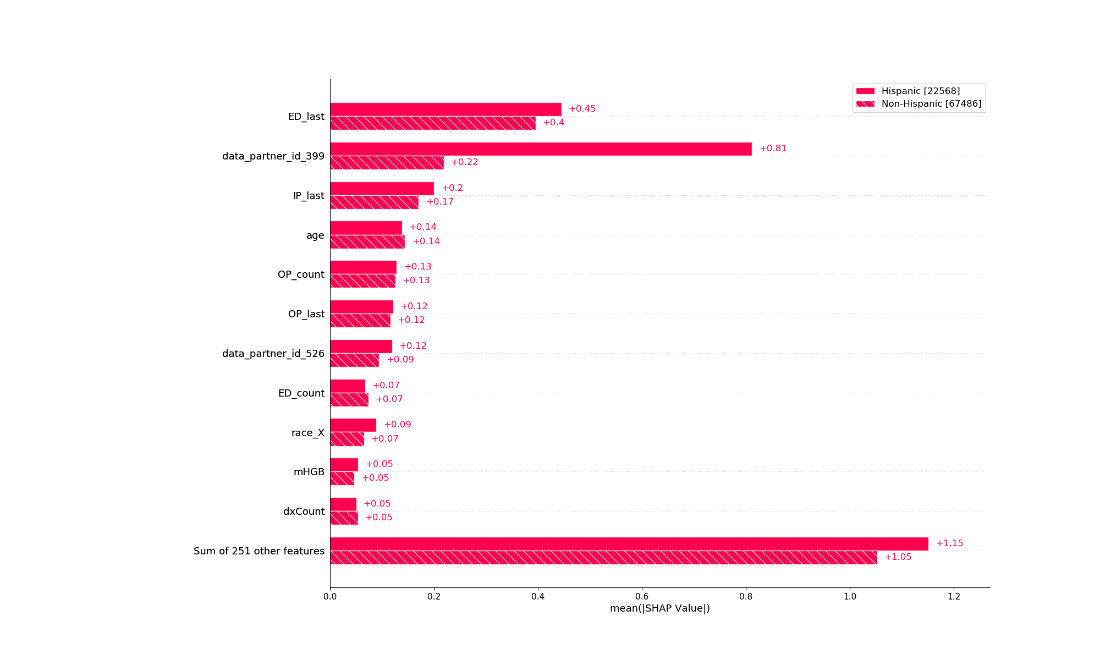


Summary of model performance for the Task #1 ensemble classifier. The classification target included all-cause hospitalization, but performed similarly for COVID and Non-COVID hospitalization with post-hoc stratification. Hosp: Hospitalization, AUCROC: Area under the receiver-operating characteristic curve, AUCPR: Area under the precision-recall curve, Acc: Accuracy, Prec: Precision.

**Figure 1: Validation set ROC and PR curves**

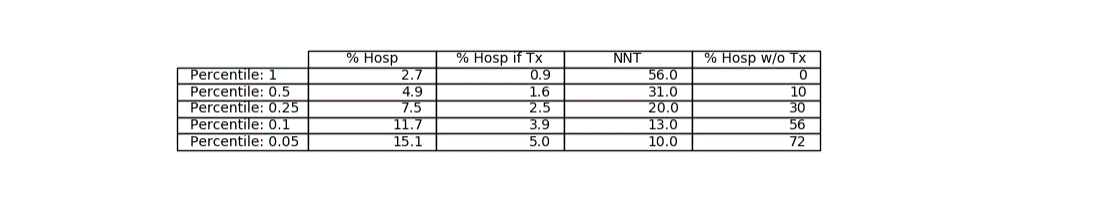


**Figure 2: Mean Shapley values stratified by ethnicity**



The top-importance predictors according to population-averaged Shapley values, stratified by ethnicity. ED/IP./OP \_last and \_count refer to the time to most recent and total count of ED, inpatient and outpatient encounters. race\_X: unknown race, mHGB: hemoglobin level, dxCount: number of total prior diagnosis codes.

**Table 3: Impact of model predictions on monoclonal antibody therapy**



Effect of targeting monoclonal antibody (mAb) therapy on the to varying percentiles of hospitalization risk. % Hosp: percent of selected patients hospitalized without mAb, % Hosp if Tx: percent of selected patients hospitalized after mAb therapy, NNT: number of children needed to treat to prevent 1 hospitalization, % Hosp w/o Tx: percentage of "missed" hospitalizations not treated with mAb

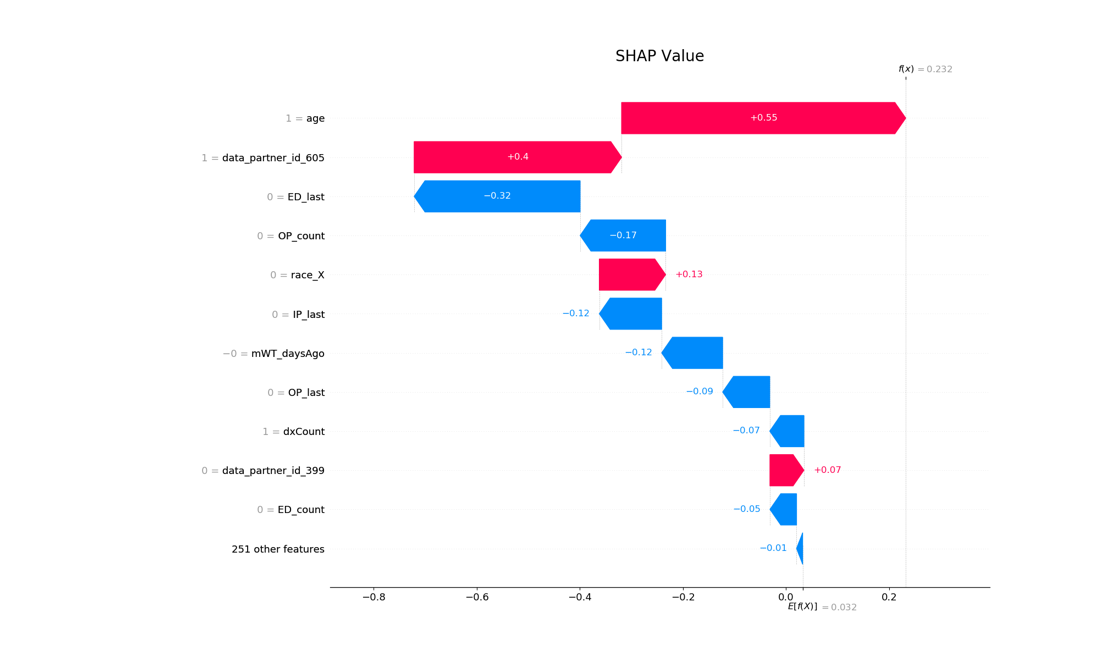
## **Conclusion/Discussion**

In the development of our ensemble classifier, we sought to provide a model which would provide clinically useful and interpretable predictions. Two of our team members are actively practicing Pediatricians (in critical care and infectious diseases). Therefore, we attempted to focus our model design and interpretation on clinical utility. Our features were designed to be easily interpreted; similar diagnosis codes and drug types were consolidated to readily-recognized clinical entities (asthma diagnosis, antibiotic use, etc.). When combined with Shapley values, we are able to provide individualized prediction explanations that reveal the person-to-person variations in influential covariates (Figure 3). Many of the highest importance variables represent patterns in past healthcare utilizations (the number and timing of past healthcare encounters), suggesting that past utilization is an important predictor of future utilization.

The significant class imbalance contributed to low precision in model predictions, however we find that in a hypothetical scenario, this model could have clinical utility. For instance, by targeting monoclonal therapy to the 25% of children with this highest risk scores, we can reduce the number needed to treat to prevent 1 hospitalization from 56 to a more clinically realistic 20, while only "missing" the opportunity to treat 30% of children eventually hospitalized.

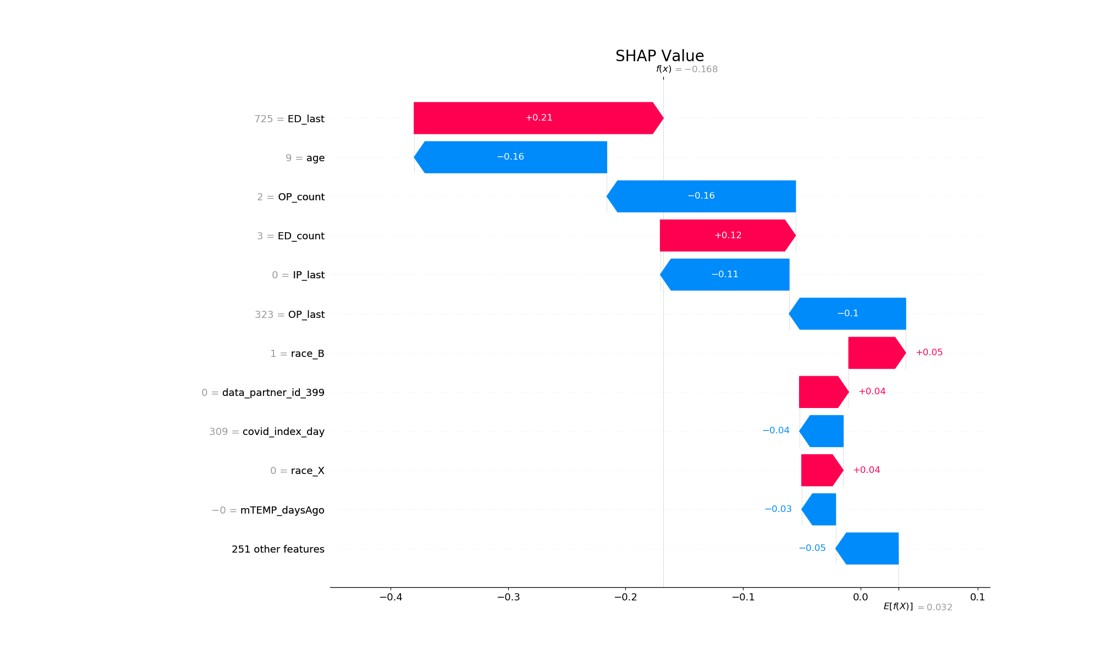
Finally, by examining stratified Shapley values, we observed that one data partner with a disproportionate number of Hispanic children contributed no hospitalizations to the data set. This may have important implications for the performance of this model on Hispanic children, as hidden racial or ethnic bias in underlying data can produce biased predictive models [7].

**Figure 3A: Shapley values for a higher-risk individual**



Additive values explaining the model prediction for a higher-risk individual. Red arrows indicate increased log-odds of hospitalization, blue arrows indicate decreased log-odds. ED/IP./OP \_last and \_count refer to the time to most recent and total count of ED, inpatient and outpatient encounters. race\_X: unknown race, mWT: weight, dxCount: count of diagnosis codes.

**Figure 3B: Shapley values for a lower-risk individual**



Additive values explaining the model prediction for a higher-risk individual. Red arrows indicate increased log-odds of hospitalization, blue arrows indicate decreased log-odds. ED/IP./OP \_last and \_count refer to the time to most recent and total count of ED, inpatient and outpatient encounters. race\_X: unknown race, race\_B: Black race, mTemp: temperature measurement.

## **References**

* Woolf SH, Chapman DA, Sabo RT, et al. Excess Deaths From COVID-19 and Other Causes, March-July 2020. JAMA. 2020 Oct 20;324(15):1562-1564.
* Kim TY, Kim EC, Agudelo AZ, Friedman L. COVID-19 hospitalization rate in children across a private hospital network in the United States: COVID-19 hospitalization rate in children. Arch Pediatr. 2021 Oct;28(7):530-532.
* Bailey LC, Razzaghi H, Burrows EK, et al. Assessment of 135 794 Pediatric Patients Tested for Severe Acute Respiratory Syndrome Coronavirus 2 Across the United States. JAMA Pediatr. 2021 Feb 1;175(2):176-184.
* Kompaniyets L, Agathis NT, Nelson JM, et al. Underlying Medical Conditions Associated With Severe COVID-19 Illness Among Children. JAMA Netw Open. 2021 Jun 1;4(6):e2111182.
* Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020 Jan;2(1):56-67.
* Weinreich DM, Sivapalasingam S, Norton T, et al. REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19. N Engl J Med. 2021 Jan 21;384(3):238-251.
* Obermeyer et al. Science. Dissecting racial bias in an algorithm used to manage the health of populations. 2019 Oct 25;366(6464):447-453.

# **Honorable Mention – Computational Methodology**

# **Multi-model Ensemble AI can Accurately Predict Pediatric COVID-19 Significant Cardiac Intervention or Invasive Ventilation Need upon Hospitalization**

Joy Alamgir

ARIScience

## **Background/Introduction**

*Hypothesis:* AI models trained on multiple clinical information views segmented by (a) all attributes (b) only current laboratory results (c) only current clinical non-laboratory presentation (d) only prior clinical presentation (e) only prior laboratory results, followed by aggregation of individual AI predicted results can result in accurate prediction of significant cardiac or invasive ventilation support need.

*Intuition behind Hypothesis:* Our intuition was to mimic access to information that clinicians may face when encountering a COVID-19 patient. Let's define INSYS for patients within the same medical information system network and OUTSYS for patients outside the same medical information system network. When a COVID-19 patient shows up at an outpatient facility or at a hospital, they may be (a) INSYS, in which case the clinician likely will have access to full patient history (b) OUTSYS, in which case the clinician will have information depending on patient's (or their caregiver's) memory of significant clinical history. In both cases the clinician has access to current laboratory results. We then extended our intuition to go a step further, and explore the possibility of predicting which patients are at most risk of serious COVID-19 *blinding* any current laboratory measurements or clinical conditions. We approached this as, should this research be successful, it will be impactful to target vaccination campaigns towards those children who are hesitant to vaccinate (on their own or because of their hesitant caregivers) but none-the-less face significant COVID-19 driven health risk. The AI models we trained correspond to these intuitive views into patient clinical information.

*Data Source:* We used BARDA provided N3C based OMOP structured data to train and validate the underlying training models

## **Methods**

*Data Preparation and Feature Pool Setup:* For Biomedical Advanced Research Development Agency (BARDA) challenge Task 2 we prepared patient data into a single master table with clinical numerical and binary attributes. Each clinical attribute then went through a dynamic selection process where the attribute's correlation with outcome of interest was compared to baseline outcome incidence rate. If the correlation was 50% higher than baseline then the attribute was eligible as a relevant feature (*"Feature Pool"*) for further assessment. A total of 298 total predictors were in the feature pool at the end of this step. The predictors used fell in the following categories and were concept set driven as per N3C/Enclave best practice:

* Demographics (age, race, gender)
* Prior need for medical care
* Charlson Comorbidity Index (divided into 6 categories: cci 0, cci 1, cci 2, cci 3, cci 4, cci >=5)
* *Prior* clinical disposition (both numerical rates and binary flags) including but not limited to defects, inflammatory markers, respiratory markers, transplants, significant medical intervention, malnutrition, cancer markers, immunocompromised markers, blood disorder markers, electrolyte imbalance markers, neurological markers, medication markers
* *Current* clinical disposition using the same markers as above at the time of *hospitalization*

The process then went into core model training as described below. 60% of the data was randomly assigned to training and 40% of the data to validation.

*Interpretability:* The features we used as mentioned in the above section are directly interpretable as each feature encapsulates a collation of related clinical states/measurements. While age, race, gender are obvious demographic features, we paid careful attention to curating concept sets that separately cluster *prior* and *current*: defects, inflammatory markers, respiratory markers, transplants, significant medical intervention, malnutrition, cancer markers, immunocompromised markers, blood disorder markers, electrolyte imbalance markers, neurological markers and medication markers. This was specifically done to establish a reasonable clinical basis behind the used features. For the reader's ease graphical depictions of highly correlated features (subset of the full feature set) are provided in Figures 1, 2, 3 and 4 below. Figure 1 focuses on pre-COVID-19 attributes, Figure 2 on binary patient condition attributes upon hospitalization, and Figure 3 on laboratory results/measurements upon hospitalization. Figure 4 shows the distribution of features per time category. A list of all features used for Task 2 is provided in Table 1 below.

*Core Model Training:* We used multiple regression and classification techniques including but not limited to a variation of subspace projection based nearest neighbor classifier that we developed based on ARIScience's prior internal work.

The *Underlying Methodologies we used for training and subsequent incorporation into decision ensembles are:*

* Decision Tree (DT)
* Random Forest (RF)
* Gradient Boosted Tree (GBT)
* Simple Linear Regression (SLR)
* Generalized Linear Regression (GLR)
* Naive Bayes Classifier (NB)
* Neural Network (NN)
* Nearest Neighbor (KNN)
* Weighted Nearest Neighbor (WKNN)
* Nearest Neighbor Subspace Projection (KNNP), and
* Weighted Nearest Neighbor Subspace Projection (WKNNP)

Each of the underlying methodologies above went through training using clinical views mentioned in the *Interpretability* sub-section above. *Additionally* each methodology went through additional training separately using features with (a) highest correlation to outcome (b) moderate correlation to outcome (c) most numerous availability - as long as such features were in the derived *Feature Pool*. The reader is reminded that a feature had to be *more* than 50% correlated as compared to *baseline outcome rate* for it to be part of the Feature Pool in the first place. For weighted variations of nearest neighbor classifiers, the outcome rate correlated with the feature served as the weight for that feature.

The total number of trained AI models were 654. Number of features used by these AI models had a mean of 173 (115). Each model achieved varying precision as shown in Figure 6 and set the basis for ensemble based predictions.

*Overfitting Mitigation:* Potential overfitting was a major concern for us given the high number of predictors and relatively low number of incidence rates resulting in imbalanced training data. Our use of multiple clinical views, along with multiple voting approaches when generating the ensemble based outcome prediction mitigates potential overfitting on any particular underlying AI model. Furthermore our selection of features into the Feature Pool itself was stringent with at-least 50% or higher outcome correlation as compared to baseline outcome rate.

*Data Imputation:* As the BARDA provided N3C data is de-identified real-world patient data, patients have widely varying completeness of data. The data imputation approach we used focused on providing neutral age-and-gender appropriate missing data for patients where data was missing with the primary aim of simulating data completeness so that the underlying AI training models can expect consistent completeness (imputed or real) for training, validation, and subsequently for BARDA-held-back test data when BARDA evaluates the submission.

*Ensemble Creation and Final Prediction:* We created 42 ensembles each with a different voting process in order to assess which ensemble performed optimally. The ensemble voting processes we incorporated are:

* At-least N votes from high precision models (N = 1, 2, 3, 4, etc)
* At-least X% vote across high and medium precision models (X = 10%, 15%, 20%)
* At-least X% vote from any AI model family (X = 10%, 20%, etc). Model families correspond to *Underlying Methodologies* section above
* At-least N votes from any high or moderate precision model (N = 1, 2, 3, etc) with voting threshold inversely correlated to precision of thee model
* At-least N votes from low precision models (N = 8, 9, 10, etc)
* Variations of each of the above with super-delegate models where a model is considered a super delegate for voting purposes if model had over X% precision (X = 65%, 70%)

At the end of ensemble creation the code automatically selects the best performing (within validation data) ensemble methodology to use for *final predictions on the BARDA held back data*.

***Summary of Novel Innovations:***Seven novel innovative approaches were undertaken beyond "typical" AI/ML techniques. These are:

* Subspace projection via collapsing features into five consistent dimensions for nearest neighbor projection searches (KNNP). This was done especially to handle widely varying levels of completeness of patient data
* Separate feature creation for both hypo and hyper (as compared to normal range) signals for laboratory results where results *below or above* reference range signify abnormal patient state
* Inclusion of prior incidence rate markers as opposed to binary markers wherever possible
* Inclusion of additional applicable hospitalizations by looking at procedure dataset in addition to visit dataset (both within the BARDA provided training data). This resulted in additional true positives and true negatives - which we believe gives our trained models an edge - as compared to BARDA provided Gold Standard (GS) files.
* Ensemble voting innovations through using super-delegates (for high performing models), scaled voting (vote weight that corresponds to model performance) and family voting concepts
* Data imputation focused on ensuring underlying classifiers can successfully process a patient record without *exception or skips,* putting in neutral signals where data was missing
* Inclusion of predictor into Feature Pool only when the predictor had 50% or more incidence rate compared to baseline incidence rate (baseline incidence rate of 11%)

*Code:* We created Java code for data preparation, (lineage color code dark purple), Java code for model training (linear color code orange) and Java code for ensemble decisions (lineage color code pink)

*Documentation:* We have followed Javadoc based documentation of the code which is the industry standard way of documenting Java based transformation including but not limited to class headings, parameter applicability, method expectations. The reader is requested to look at the following two representative Java code files in relation to documentation standard example of the challenge Java files: [Java File Example 1](https://unite.nih.gov/workspace/data-integration/code/repos/ri.stemma.main.repository.31981fab-c05a-4b78-925d-4ea83109bf97/contents/refs%2Fheads%2Fmaster/transforms-java/src/main/java/myproject/datasets/MasterTableSupport1c_covid.java) and [Java File Example 2](https://unite.nih.gov/workspace/data-integration/code/repos/ri.stemma.main.repository.31981fab-c05a-4b78-925d-4ea83109bf97/contents/refs%2Fheads%2Fmaster/transforms-java/src/main/java/myproject/datasets/MasterTable.java)

## **Conclusion/Discussion**

The ensemble of ensembles AI process we setup allows prediction of significant cardiac/invasive ventilation intervention outcomes. Since we do not have a-priori access to BARDA held-back data we can only report on our ensemble model performance on validation data in this writeup. As such, following is a snapshot view of Task 2 ensemble performance on validation data.

*Validation Data Performance Summary:* TPR: 85.5%, TNR: 95.3% FPR: 4.7%, AUROC: 0.926, F2: 0.813

*Key Takeaways:*

* Prediction of significant cardiac or ventilation support need can be made for a subset of patients even ***without*** hospitalization - based purely on historical patient clinical attributes. The reason this finding is significant is that targeted vaccination campaigns for vaccination hesitant children who none-the-less are at high risk can reduce hospitalization burden and severe health outcome burden within the US - and this particular prediction subset helps identify that target accurately (TPR: 72.9%, TNR: 97.7% FPR: 2.4%, AUROC: 0.85). As such this particular AI model subset has immediate applicability in the COVID-19 pandemic.
* Ensemble AI is pivotal for predicting COVID-19 pediatric outcomes due to imbalanced data and to reduce overfitting on any specific model
* Post-hospitalization significant cardiac or invasive ventilation support prediction can be readily encapsulated through stateless, session-less JSON services for rapid use across the US as clinical information to providers and acute care facilities. For facilities that already use or provide that in OMOP the predictions can be more natively incorporated. Predictions can also be fed back to the data providers providing data to N3C.
* Post-hospitalization significant cardiac or invasive ventilation support prediction is comprehensive
* The AI approach formulated by the authors can be repeated by the authors for other non-COVID-19 binary outcome prediction using OMOP structured data
* Predictions are consistent / pseudo-deterministic - particularly as the use of "ensemble of ensembles" smooth out variations in inherently random (e.g. one of the many underlying models we used is RandomForest) training processes. The implementation is structured to work with varying levels of OMOP inputs and compute environments.

*Future beyond COVID-19:* We have structured the code so that substantial portions of it can be *re-used* by the authors for binary outcome AI projects using OMOP/N3C structured data. While pediatric post-hospitalization cardiac/invasive ventilation intervention is today's AI goal, tomorrow it can be *re-purposed* by the authors for other outcomes or other diseases. Furthermore the concept sets we have used to setup features have been defined with repurpose goals in mind.

*Pitfalls:* Setting up the concept sets that drove both prior and current feature establishment for patients took substantial curation - particularly in minimizing overlaps between concept sets so as to minimize overlapping signals for the AI models. An immediate next step would be to further refine the concept sets used to reduce overlapping signals.

## **References**

* UCSF Benioff Children's Hospital. Accessed: Oct 31 2021. Updated: Continually. <https://www.ucsfbenioffchildrens.org/medical-tests>
* Merck Manuals. Accessed Oct 31 2021. Updated Continually. <https://www.merckmanuals.com/professional/hematology-and-oncology>
* Mayo Clinic Laboratories Pediatric Test Catalog. Accessed Oct 31 2021. Updated: Continually. <https://pediatric.testcatalog.org>
* Martin B, DeWitt PE, Russell S, Anand A, Bradwell KR, Bremer C, Gabriel D, Girvin AT, Hajagos JG, McMurry JA, Neumann AJ, Pfaff ER, Walden A, Wooldridge JT, Yoo YJ, Saltz J, Gersing KR, Chute CG, Haendel MA, Moffitt R, Bennett TD. Children with SARS-CoV-2 in the National COVID Cohort Collaborative (N3C). medRxiv [Preprint]. 2021 Jul 23:2021.07.19.21260767. doi: 10.1101/2021.07.19.21260767. PMID: 34341796; PMCID: PMC8328064.
* Dionne A, Kheir JN, Sleeper LA, Esch JJ, Breitbart RE. Value of Troponin Testing for Detection of Heart Disease in Previously Healthy Children. J Am Heart Assoc. 2020 Feb 18;9(4):e012897. doi: 10.1161/JAHA.119.012897. Epub 2020 Feb 13. PMID: 32067577; PMCID: PMC7070204.
* Uddin S, Khan A, et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):281. Published 2019 Dec 21. doi:10.1186/s12911-019-1004-8
* US CDC Growth Charts. Accessed Oct 31 2021. Updated: Dec 7 2016. <https://www.cdc.gov/growthcharts/cdc_charts.htm>
* Melissa A Haendel, Christopher G Chute, Tellen D Bennett, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment, JAMIA Aug 2020. DOI: 10.1093/jamia/ocaa196
* Wikipedia. F-Measure, TPR, TNR, FPR, FNR. Accessed Oct 20,2021. Updated Continually. <https://en.wikipedia.org/wiki/F-score>
* BARDA Community Challenge - Pediatric COVID-19 Data Challenge ([syn25875374](https://www.synapse.org/#!Synapse:syn25875374/wiki/611225)).