**Clinical predictors of COVID-19 mortality:**

The scripts in this repository is used to reproduce the results for “clinical predictors of COVID-19 mortality”. Program files are shared in the folders by figures name (Figure 2, Figure 3, and Figure 4). In folder Figure 1.

**Setup environments:** This project is developed in Jupyter Notebook environment. So the following are required:

* Python 3.7.3
* Pandas 0.24.2
  + NumPy 1.16.2
* Scikit-learn 0.20.3
* Scipy 1.2.1
* Matplotlib 3.1.1

**Data:**

Anonymized electronic medical record (EMR) data from patients diagnosed with COVID-19 within the Mount Sinai Hospital System, New York, NY from March 9th through April 6th, 2020 were included in the study. On 6th April, 5051 COVID-19 positive patients treated at the Mount Sinai Health System, data split in training set and test set.

**Missing Value Imputation:** In figure 2A, we have We first attempted to find the optimal percentage of missing values in each variable across the patients in the development set (missing value level) that could be imputed and lead to more accurate prediction. In this script, we have pre-processed the development set data and split it randomly 100 times into training and validation set. Used four different classifiers (Random Forest, Logistic Regression, Support Vector Machine and XGBoost) with increment of 5% missing value imputation.

**RFE model:** In figure 2B, we used a setup analogous to missing value imputation, and the Recursive Feature Elimination (RFE) algorithm, we evaluated the performance of the four classification algorithms with different number of features selected from the full set of features. The average AUC scores from 100 runs of this process are shown here, along with error bars.

Figure 3, script is self-descriptive. Here, we have plotted AUC curve based on optimum imputation features (17F) model and top three features from RFE 3F model. Also calibration of the plot has been done by using ‘calibration\_curve’ library from sklearn.

Figure 4, script is self-descriptive, here we have ran RFE simulation simulations for top three features. This script generate three independent hundred simulation of RFE model for four classifiers by using 17 features and selected top three features and saved in a csv file and plot horizontal bar graphs. One can develop figure 4A of the paper by choosing single seed of 300 runs or by summation of three runs output data.

Characteristic table p-value and odd ratio calculation is provided by (‘table\_pval\_effect\_size.py’), we used Cohen’s\_d to determine the effect size of the experiment and ttest to calculate p-value.

To site this paper (**doi:** <https://doi.org/10.1101/2020.05.19.20103036>)