Developing a Machine Learning Framework for Predicting Severe COVID-19 Disease and Investigating Risk Factors in Vaccinated and High-Risk Populations

Kanika Mahajan[1], Alex Bachmann[2], Richard Beswick[2], Xifeng Yan[1]
[1]Department of Computer Science, University of California Santa Barbara
[2]The Cottage Health Research Institute, Santa Barbara, California
Contact e-mail: kanikamahajan79@gmail.com, xyan@ucsb.edu

1. INTRODUCTION

COVID-19 has affected millions across the globe and continues to be a significant consideration in the foreseeable future [1-4]. Even though several vaccines have emerged and proven to be effective, COVID-19 is far from over given the appearance of newer and more virulent strains of the novel SARS- CoV-2 [4,5,6]. Despite strict control measures implemented worldwide and increase in vaccinations globally, several reports indicate that various variants have enhanced transmission and infectivity while reducing recognition by host antibodies [6].

COVID-19 disease has some major concerns. First, it shows sudden progression to critical illness and high mortality in some patients [7-10]. Rapid and effective triage strategies are critical for optimal treatment and appropriate allocation of hospital resources [7,8,11]. Also, early identification of such sub-groups of COVID-19 cases that are at risk of progressing to severe infection is important for precise, timely, and targeted treatment delivery [7,10,14].

Several clinical prognostic models have been published previously to predict adverse prognostic outcomes and clinical deterioration among COVID-19 patients presenting to hospitals. However, it is known that these models suffer from various problems such as lack of generalizability due to small sample-size, high risk of bias, overestimation of prediction accuracy, and lack of clinical utility as models are not based on easily available patient data-points [10,11,12,13,14].

Second, Covid-19 has proven to be more severe and deadly than the seasonal flu viruses and disproportionately affects some demographics and subpopulations [1,2,3]. It remains unclear whether the higher rates of severe disease and deaths observed globally among minority ethnic

groups are attributed to an increased risk of infection, a worse prognosis, or a combination of both factors [1,2,15,16].

Our research seeks to address these challenges and pinpoints clinical risk factors associated with the progression to severe COVID-19 across the broader population, while also going deeper into high-risk sub-populations. Given substantial research exists our main value addition with this study is twofold. First, we are incorporating COVID-19 vaccination data into our analysis of individuals who were diagnosed with COVID-19 during the early stages of the pandemic. Our algorithms will utilize existing clinical data normally collected for COVID-19 patients (like demographics, medical history, signs, and symptoms etc.) along with patients' vaccination status to capture early warning signals of the need for more aggressive treatment in the overall study population and those who are already vaccinated with covid. This will help in estimating the probability of prognostic clinical outcomes to assist in preparation of suitable treatment plans.

Second, to address open questions in the literature on risk factors associated with severe covid for certain subpopulations we identify risk factors associated with severe covid in 2 subpopulations: those with history of diabetes and individuals with hispanic origin to get a deeper understanding of what drives COVID in these high risk groups.

We believe our research will help to a) improve our understanding of the virus and its evolving effects, b) identify positive cases that escape the effects of COVID vaccines, and have a propensity to progress to a severe disease c) provide a deeper understanding of risk factors in high risk populations such as Hispanic population and those with pre-existing diabetes.

2. MATERIALS AND METHODS

2.1 Study Population

The eligible population for this study was COVID-19 positive patients presenting to Santa Barbara Cottage Hospital between March 2020 and April 2022. Inclusion criteria was patients >= 18 years of age that were diagnosed with COVID-19 confirmed through standard COVID-19 testing methods. There were multiple records for some patients, but data used for this study is the data available during initial presentation and same day inpatient admission. Any variables with >50% of the data missing were excluded from the analysis. A small subset of patients had multiple admissions during the entire study time frame at cottage hospital. As this is not a longitudinal study, for such patients we used the visit which had the maximum duration of hospital stay to capture more severe cases in this study. A total of 2940 unique patients were enrolled in this study out of which 2534 were labeled as non-severe and 406 (~14%) were labeled as severe cases. Patients were labeled as severe if they met any of the following criteria:

- received invasive respiratory treatment like mechanical ventilation
- event of ICU admission
- length of stay in hospital was > 7 days
- death at discharge

2.2 Data Collection and Study Design

This is a retrospective cohort study and data used in the study was previously recorded in a natural setting. This negates the possibility of selection and information bias affecting this study results. Data was collected using manual and automated abstraction methods from patient electronic medical records from the Santa Barbara Cottage Health System (CottageOne).

2.3 Analysis

2.3.a. Descriptives and univariate tests

We conducted descriptive statistical analysis to describe this study population and their attributes (continuous and categorical) that were available to build these models. The variables included in this study can be broadly classified into demographics, history of hospital stay, covid vaccination history, comorbidities, vital signs and symptoms, presence of lung abnormalities (ground-glass opacities or bilateral consolidation in the peripheral lower lung fields) and pulmonary fibrosis on radiological exam, re-infection, and history of tobacco use. Laboratory data was not included due to high missing rates.

We also conducted univariate tests to assess the association of each predictor with severe COVID-19. All analyses were performed in Python V.3. Two independent samples were tested by the Wilcoxon rank sum test. The $\chi 2$ (Chi-square) test of independence was performed to compare count data, and a 2-tailed value of $P < 0.05$ was considered statistically significant throughout the study. On analysis except history of Gout, liver disease, Hepatitis, presence of pulmonary fibrosis and reinfection, all other features were statistically significant on univariate analysis. This was a strong foundation for this study as the majority features are likely to offer valuable predictive information regarding the outcome variable (i.e. severe covid) critical for developing robust models.

For reference detailed results are available in the Appendix section 1.

2.3.b. Feature selection:

For feature selection first we ran univariate tests on each predictor and only included those which had statistically significant association with our target variable i.e. severe covid disease. Features like SBP and DBP, RR and SpO2 are known to be prone to multicollinearity. In order to avoid this issue and still retain these important features we created 2 interaction terms. First was the

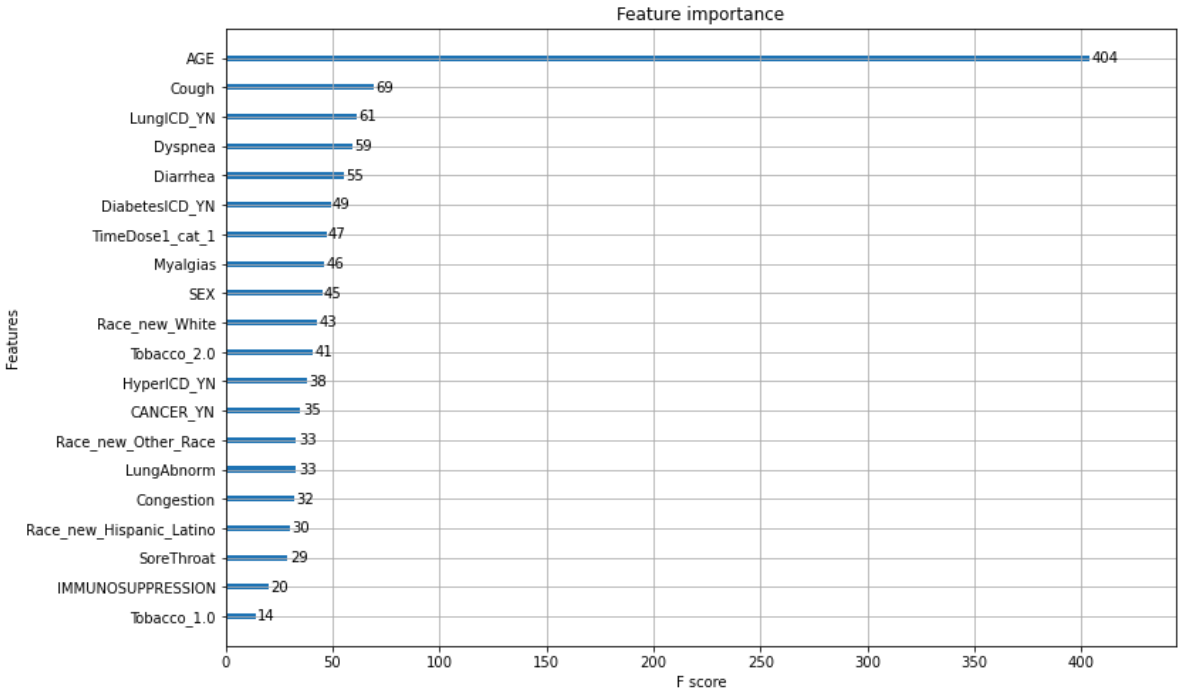
Mean arterial pressure calculated as DBP + pulse pressure/3 and pulse_pressure calculated as SBP-DBP. Second interaction term was simply SpO2 * RR. For each of our models we checked for multicollinearity using VIF criterion and any features with a VIF > 15 were excluded from model building. For subset analysis features that could not be inverted and led to a singular matrix were also excluded.

2.3.c. Models:

2.3.c.a Severe COVID prediction in patients at hospital admission including patient's vaccination status:

Our first model was built to predict severe disease in the overall study population at the time of hospital admission. Oversampling of minority class was performed to correct for imbalance. We used the XgBoost algorithm for model construction, utilizing grid search and 5-fold cross-validation for hyperparameter optimization. The parameters of our best fitting model were: 'alpha': 1.0, 'colsample_bytree': 0.9, 'gamma': 2.0, 'lambda': 2.0, 'learning_rate': 0.2, 'max_depth': 4, 'n_estimators': 300, 'subsample': 0.9. Overall model accuracy was 87%, ROC AUC was 0.91 and macro precision was 0.72, macro avg recall was 0.83 and macro-average F1 score was 0.75. ***The top 5 features increasing the risk for severe covid based based on shapley analysis were increasing Age, less cough, history of Lung ICD disease, symptoms of dyspnea and diarrhea.*** The RoC curve and Shapley analysis graphs can be found in Appendix section 2.

	precision	recall	f1-score	support
0	0.97	0.88	0.92	519
1	0.46	0.77	0.58	69
accuracy			0.87	588
macro avg	0.72	0.83	0.75	588
weighted avg	0.91	0.87	0.88	588



2.3.c.b Risk factors of severe COVID in vaccinated individuals:

Our next model was to identify risk factors associated with severe COVID in individuals with at least 1 or more doses of COVID vaccines before they were COVID positive. There were only 825 patients identified as COVID vaccinated and out of these only 49 had severe COVID. Given the small sample size and severe class imbalance in our training data we decided not to use ML algorithms to avoid overfitting our model. Instead we used the Newton Method for Logistic regression that is robust to small sample sizes to understand the log odds of severe COVID infection as explained by the features that we had. Our model had a decent pseudo R2 of 39% and was overall statistically significant with a p value< 0.05. **Coefficients for Age, Lung abnormalities, history of Diabetes were all statistically significant and positive. Out of these, radiological presence of lung abnormalities had the highest coefficient 1.7 (odds ratio = 5.7) which means that the odds of severe covid increases ~6 times in those with lung abnormalities during initial presentation to the hospital compared to those with no abnormalities.**

Logit Regression Results

Dep. Variable:	Severe	No. Observations:	640
Model:	Logit	Df Residuals:	622
Method:	MLE	Df Model:	17
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.3922
Time:	01:00:30	Log-Likelihood:	-68.008
converged:	True	LL-Null:	-111.89
Covariance Type:	nonrobust	LLR p-value:	1.547e-11

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.3569	1.586	-4.638	0.000	-10.466	-4.248
AGE	0.0395	0.020	1.971	0.049	0.000	0.079
SEX_1	0.2013	0.514	0.391	0.696	-0.807	1.209
IMMUNOSUPPRESSION_1	1.3039	0.810	1.610	0.107	-0.283	2.891
LungAbnorm_1	1.7374	0.597	2.910	0.004	0.567	2.908
CANCER_YN_1	0.0185	0.607	0.030	0.976	-1.172	1.209
Cough_1	-0.8354	0.510	-1.638	0.101	-1.835	0.164
WetCough_1	0.9343	0.849	1.100	0.271	-0.730	2.598
Diarrhea_1	0.5618	0.604	0.930	0.353	-0.623	1.746
Dyspnea_1	-0.2987	0.574	-0.520	0.603	-1.424	0.826
Myalgias_1	-0.6658	0.501	-1.328	0.184	-1.648	0.317
Congestion_1	-1.4906	1.166	-1.278	0.201	-3.777	0.795
SoreThroat_1	-0.0707	0.852	-0.083	0.934	-1.740	1.598
HyperICD_YN_1	0.3840	0.596	0.644	0.520	-0.784	1.552
LungICD_YN_1	1.0997	0.569	1.933	0.053	-0.015	2.215
DiabetesICD_YN_1	1.2902	0.515	2.504	0.012	0.280	2.300
Race_new_White_1	-0.1016	0.614	-0.165	0.869	-1.304	1.101
Tobacco_0.0_1	0.6094	0.523	1.164	0.244	-0.416	1.635

2.3.c.c Risk Factors of severe COVID in Hispanic Population:

Our next subpopulation was the Hispanic population who are known to suffer from high risk of severe covid. To understand this better in our next model we identified risk factors that are associated with severe COVID in hispanic population. There were only 201 patients identified as Hispanics and out of these only 51 had severe COVID. Again given the small sample size we used the Newton Method for Logistic regression. Our model had a decent pseudo R2 of 39% and was overall statistically significant with a p value< 0.05. **Coefficients for Age, symptoms of Cough and Congestion were statistically significant. Out of these, symptoms of congestion had the highest positive coefficient of 2.3 (odds ratio = 10.3) which means the odds of severe covid increases ~10 times in hispanic population with congestion during initial presentation to the hospital compared to those with no abnormalities.**

On the other hand, cough had the highest negative coefficient of ~-1.5 (odds ratio = 0.22) which means that the odds of having severe COVID decreased by 78% (1-0.22) in our hispanic study population with cough compared to those that did not have cough.

Logit Regression Results

Dep. Variable:	Severe	No. Observations:	142
Model:	Logit	Df Residuals:	126
Method:	MLE	Df Model:	15
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.3853
Time:	01:03:10	Log-Likelihood:	-45.005
converged:	True	LL-Null:	-73.220
Covariance Type:	nonrobust	LLR p-value:	1.025e-06

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.6188	1.752	-3.207	0.001	-9.053	-2.184
AGE	0.0680	0.024	2.809	0.005	0.021	0.115
SEX_1	0.5324	0.652	0.816	0.414	-0.746	1.811
LungAbnorm_1	0.3071	0.921	0.333	0.739	-1.498	2.112
CANCER_YN_1	-0.6380	0.988	-0.646	0.518	-2.574	1.298
Cough_1	-1.4706	0.656	-2.241	0.025	-2.756	-0.185
WetCough_1	2.9934	1.696	1.765	0.078	-0.331	6.317
Diarrhea_1	0.5403	0.689	0.785	0.433	-0.809	1.890
Dyspnea_1	1.0624	0.711	1.494	0.135	-0.331	2.456
Myalgias_1	-0.7295	0.568	-1.285	0.199	-1.842	0.383
Congestion_1	2.3318	0.906	2.574	0.010	0.556	4.107
SoreThroat_1	-0.6614	0.996	-0.664	0.507	-2.614	1.292
HyperICD_YN_1	0.3528	0.661	0.534	0.594	-0.943	1.649
LungICD_YN_1	1.2183	0.773	1.576	0.115	-0.297	2.733
DiabetesICD_YN_1	0.0239	0.687	0.035	0.972	-1.322	1.370
Tobacco_0.0_1	-0.6955	0.766	-0.908	0.364	-2.196	0.805

2.3.c.d Risk Factors of severe COVID in Diabetic Population:

As per literature search Diabetes population seemed to suffer from high risk of severe covid. To understand this better in our next model was to identify risk factors associated with severe COVID in diabetic population. There were 493 patients identified as diabetics and out of these only 124 had severe COVID. Again given the small sample size we used the Newton Method for Logistic regression. Our model had a decent pseudo R2 of 35% and was overall statistically significant with a p value< 0.05. **Coefficients for Age, presence of lung abnormalities, symptoms of wet cough and myalgia, and history of lung disease were statistically significant. Out of these, history of lung disease had the highest positive coefficient of 1.9 (odds ratio = 6.4) which means that the odds of having severe COVID is ~6 times more in diabetics with a history of lung disease compared to those who did not.**

On the other hand, myalgia had the highest negative coefficient of -0.8 (odds ratio = 0.44) which means that the the odds of having severe COVID decreased by 56% (1-0.44) in our diabetic study population with myalgia compared to those that did not have myalgia*.

Logit Regression Results

Dep. Variable:	Severe	No. Observations:	448
Model:	Logit	Df Residuals:	434
Method:	MLE	Df Model:	13
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.3468
Time:	01:04:38	Log-Likelihood:	-162.37
converged:	True	LL-Null:	-248.58
Covariance Type:	nonrobust	LLR p-value:	5.958e-30

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.4429	0.910	-5.983	0.000	-7.226	-3.660
AGE	0.0395	0.012	3.313	0.001	0.016	0.063
SEX_1	0.1026	0.282	0.364	0.716	-0.450	0.655
LungAbnorm_1	1.4322	0.412	3.476	0.001	0.625	2.240
CANCER_YN_1	-0.8470	0.468	-1.811	0.070	-1.763	0.069
Cough_1	-0.4616	0.300	-1.540	0.124	-1.049	0.126
WetCough_1	1.0664	0.526	2.027	0.043	0.035	2.097
Diarrhea_1	-0.1237	0.374	-0.331	0.741	-0.857	0.610
Dyspnea_1	0.3914	0.310	1.263	0.207	-0.216	0.999
Myalgias_1	-0.8301	0.284	-2.923	0.003	-1.387	-0.273
Congestion_1	-0.3250	0.682	-0.476	0.634	-1.662	1.012
SoreThroat_1	0.0670	0.508	0.132	0.895	-0.928	1.062
HyperICD_YN_1	0.1306	0.323	0.404	0.686	-0.503	0.764
LungICD_YN_1	1.8578	0.357	5.198	0.000	1.157	2.558

* Please note myalgia in our study is capturing all aches except chest pain and ear aches, headaches and weakness.

3. DISCUSSION

3.1 Findings

From the findings in our study we had some interesting observations and some results that confirm our existing knowledge about predictors of severe COVID. As well known age, history of diabetes, lung disease emerged among top 6 predictors of severe COVID. Symptom data: dyspnea, diarrhea and cough also emerged to be powerful predictors of severe covid and in top 6.

In our subpopulation analysis Age was a predictor and risk factor in all models as we would have expected. History of lung disease or presence of lung abnormalities on radiological examination also appeared to be common risk factors across all sub-populations except hispanic population.

However, it was interesting to note that symptoms exhibited distinct behavior in our different sub-populations. Symptoms were not a risk factor in covid vaccinated individuals but had statistically significant association with severe covid in both our high-risk populations: hispanics and those with pre-existing diabetes. It's further interesting to note that the symptoms that were risk factors also differ between both sub-populations. Cough and congestion were significant in

hispanics, and wet cough and myalgia in diabetic population. More research and analysis would be beneficial to understand this further and hypothesis that could explain these associations.

3.2 Limitations

Overall our study had some limitations. First, we have not included COVID-19 variant information in our analysis as this was not available in our data. However, given the time frame of our study is March 2020 - April 2022 our population would mostly be infected by pre-omnicorn era variants (omicron variant emerged in November 2021). This means bias that could arise due to comparing study participants infected with more severe strains is low.

Second, for our subpopulation analysis (those with history of covid vaccine, hispanics and diabetes) the sample size was less to build machine learning models. Instead we ran statistical models which are known to be robust to small sample sizes to understand risk factors associated with severe COVID in these populations.

Third, our study includes data from one hospital but our study center is the largest in the Central California coast and does service a large geographical region. However if we were to include more hospitals from different regions we could increase the generalizability of this study. Last, we were not able to include any laboratory variables in analysis due to high missing rates.

3.3 Conclusion and future research

The COVID-19 pandemic stands as the most severe health crisis to affect humanity in recent times. Managing the pandemic has been extremely challenging, and the emergence and rapid spread of Variants make this even more demanding. The pandemic has already caused significant economic and social hardships, claiming millions of lives worldwide. The emergence and spread of viral variants raise concerns that our global fight against the pandemic will extend far longer than anticipated.

The application of machine learning (ML) algorithms for predicting COVID-19 prognosis shows substantial potential. Broad implementation of these algorithms could significantly mitigate the strain on healthcare resources and improve patient outcomes[13].

The urgent question we face is not if there will be a future pandemic or epidemic but actually when will the next health emergency arise. In this unpredictable landscape a primary concern is ensuring accurate prevention, preparedness, and prediction for future pandemics[14]. We hope from insights provided from this study we were able to contribute towards better preparedness towards future pandemics by empowering medical providers with evidence-based triage strategies.

REFERENCES

1. Irizar, P., Pan, D., Kapadia, D., Bécares, L., Sze, S., Taylor, H., ... & Pareek, M. (2023). Ethnic inequalities in COVID-19 infection, hospitalisation, intensive care admission, and death: a global systematic review and meta-analysis of over 200 million study participants. *EClinicalMedicine*, *57*.
2. Pan, D., Sze, S., Martin, C. A., Nazareth, J., Woolf, K., Baggaley, R. F., ... & Pareek, M. (2021). Covid-19 and ethnicity: we must seek to understand the drivers of higher transmission. *bmj*, *375*.
3. Kim, M. I., & Lee, C. (2023). Human coronavirus OC43 as a low-risk model to study COVID-19. *Viruses*, *15*(2), 578.
4. Del Rio, C., & Malani, P. N. (2023). COVID-19 in the Fall of 2023—Forgotten but Not Gone. *JAMA*.
5. Koff, W. C., & Berkley, S. F. (2021). A universal coronavirus vaccine.
6. Gong, W., Parkkila, S., Wu, X., & Aspatwar, A. (2023). SARS-CoV-2 variants and COVID-19 vaccines: Current challenges and future strategies. *International reviews of immunology*, *42*(6), 393-414.
7. Kamal, M., Hasan, S. T., Sarmin, M., Das, S., Shahrin, L., Faruque, A. S. G., ... & Ahmed, T. (2024). Prognostic accuracy of early warning scores for predicting serious illness and in-hospital mortality in patients with COVID-19. *PLOS Global Public Health*, *4*(3), e0002438.
8. Liang, W., Yao, J., Chen, A., Lv, Q., Zanin, M., Liu, J., ... & He, J. (2020). Early triage of critically ill COVID-19 patients using deep learning. *Nature communications*, *11*(1), 1-7.
9. Sun, C., Hong, S., Song, M., Li, H., & Wang, Z. (2021). Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning. *BMC Medical Informatics and Decision Making*, *21*(1), 1-16.
10. Gupta, R. K., Harrison, E. M., Ho, A., Docherty, A. B., Knight, S. R., van Smeden, M., ... & Metelmann, S. (2021). Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study. *The Lancet Respiratory Medicine*, *9*(4), 349-359.
11. Gao, Y., Cai, G. Y., Fang, W., Li, H. Y., Wang, S. Y., Chen, L., ... & Gao, Q. L. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature communications*, *11*(1), 1-10.
12. Cecconi, M., Piovani, D., Brunetta, E., Aghemo, A., Greco, M., Ciccarelli, M., ... & Bonovas, S. (2020). Early predictors of clinical deterioration in a cohort of 239 patients hospitalized for Covid-19 infection in Lombardy, Italy. *Journal of clinical medicine*, *9*(5), 1548.
13. Chen, R., Chen, J., Yang, S., Luo, S., Xiao, Z., Lu, L., ... & Xu, J. (2023). Prediction of prognosis in COVID-19 patients using machine learning: a systematic review and meta-analysis. International Journal of Medical Informatics, 105151.

14. Coccia, M. (2023). Sources, diffusion and prediction in COVID-19 pandemic: lessons learned to face next health emergency. *AIMS Public Health*, *10*(1), 145.
15. Floyd, J. S., Walker, R. L., Kuntz, J. L., Shortreed, S. M., Fortmann, S. P., Bayliss, E. A., ... & Dublin, S. (2023). Association between diabetes severity and risks of COVID-19 infection and outcomes. *Journal of General Internal Medicine*, *38*(6), 1484-1492.
16. Li, R., Shen, M., Yang, Q., Fairley, C. K., Chai, Z., McIntyre, R., ... & Zhang, L. (2023). Global diabetes prevalence in COVID-19 patients and contribution to COVID-19–related severity and mortality: a systematic review and meta-analysis. *Diabetes Care*, *46*(4), 890-897.

Appendix

Appendix 1.

Descriptive statistics by severity along with tests for difference in means and medians

Variable	**Mean (std.)**		**Median**		***P-value****		***Missing***
	Severe	Non-Severe	Severe	Non-Severe	t-test	*Wilcoxon rank sum*	
Age	66	50	69	49	<0.01	<0.01	0%
LOS	16	1	11	0	<0.01	<0.01	0%
ICU LOS	4	0	0	0	<0.01	<0.01	0%
Fever	98.2	98	98	97.8	0.048	<0.01	3.2%
Pulse	96	91	94	90	<0.01	<0.01	1.7%
RR	24	19	22	18	<0.01	<0.01	2.6%
SpO2	90	97	93	98	<0.01	<0.01	1.0%

*t-test with Wald's test for unequal variances using 2 sided confidence level

Demographics and Hospital Stay	**Non-Severe**	**Severe**	**Missing (%)**	***P-value****
Race			0%	<0.01
White	1740	266		
Other_Race	554	71		

Hispanic_Latino	155	51		
African_American	55	10		
Unknown	30	8		
Gender			0%	<0.01
Male	1270	243		
Female	1264	163		
ICU Stay			0%	<0.01
No	2534	204		
Yes	0	202		
Invasive Ventilation			0%	<0.01
No	2534	319		
Yes	0	87		
Death at Discharge			0%	<0.01
No	2534	314		
Yes	0	92		
Covid Vaccine Dose Status				<0.01
No date of dose information (2)	1108	216	0%	
Vaccine received before positive test (1)	776	49		
Vaccine received after positive test (0)	650	141		
COMORBIDITIES				
Variable	Non-Severe	Severe	Missing (%)	*P-value**
Diabetes_YN			0%	<0.01
0	2126	235		
1	407	171		
DiabetesICD_YN			0%	<0.01
0	2165	282		

1	369	124		
HyperICD_YN			0%	<0.01
0	1919	250		
1	615	156		
GoutICD_YN			0%	0.10
0	2519	400		
1	15	6		
LungICD_YN			0%	<0.01
0	2122	188		
1	412	218		
IMMUNOSUPPRESSION			0%	<0.01
N	2498	381		
Y	36	25		
LIVER_DISEASE			0%	0.84
N	2461	393		
Y	73	13		
HEPATITIS-B_YN			0%	0.07
N	2533	404		
Y	1	2		
CANCER_YN			0%	<0.01
N	2341	357		
Y	193	49		
Tobacco			513 (17.4%)	<0.01
Never Used(0)	1663	157		
Current User(1)	207	11		
Former User(2)	328	61		
Signs and Symptoms				

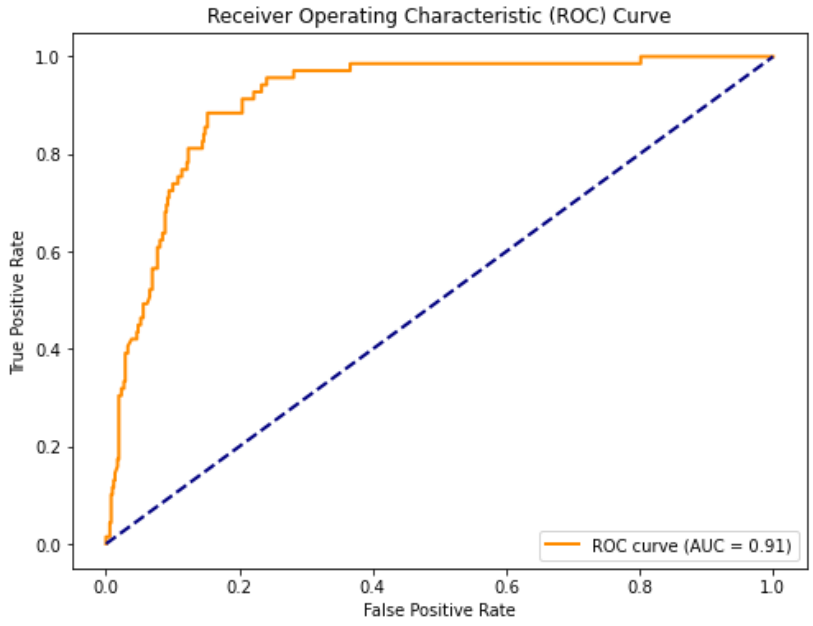
Variable	Non-Severe	Severe	Missing (%)	*P-value**
Cough			12%	<0.01
1	839	126	208	
0	1487	131	149	
WetCough			12%	<0.01
0	2243	233	210	
1	81	24	149	
Diarrhea			12%	<0.01
0	1958	213	208	
1	368	44	149	
Dyspnea			12%	<0.01
0	1561	87	211	
1	762	170	149	
Myalgias			12%	<0.01
1	966	153	208	
0	1360	104	148	
Wheeze			12%	<0.01
0	2270	255	208	
1	56	2	149	
Congestion			12%	<0.01
0	1974	245	208	
1	352	11	150	
SoreThroat			12%	<0.01
0	1671	241	208	
1	655	16	149	
Hemoptysis			12%	<0.01
0	2305	250	210	

1	19	7	149	
PulmFibrosis_YN				0.21
0	2520	401	0%	
1	14	5		
LungAbnorm			12.4%	<0.01
0	1787	26	217	
1	530	231	149	
Reinf_YN				0.81
0	2522	405	0%	
1	12	1		

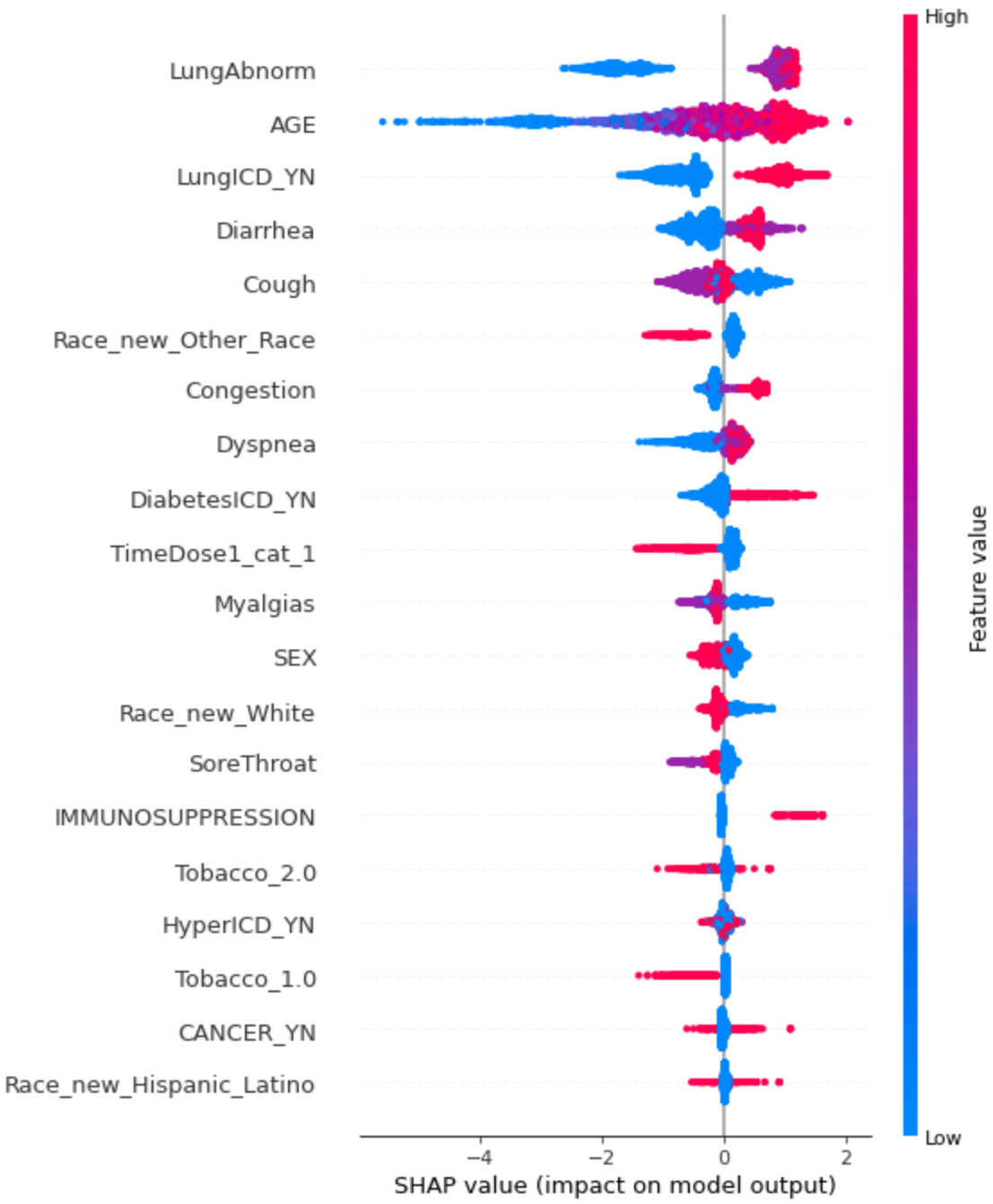
Categorical variables (Total should be 2940 for all)*Chi square test with 2 sided confidence level

Appendix 2.

ROC CURVE



SHAPLEY ANALYSIS


LungAbnorm
AGE
LungICD_YN
Diarrhea
Cough
Race_new_Other_Race
Congestion
Dyspnea
DiabetesICD_YN
TimeDose1_cat_1
Myalgias
SEX
Race_new_White
SoreThroat
IMMUNOSUPPRESSION
Tobacco_2.0
HyperICD_YN
Tobacco_1.0
CANCER_YN
Race_new_Hispanic_Latino
−4
−2
0
2
SHAP value (impact on model output)
High
Low
Feature value