MACHINE LEARNING ANALYSIS OF DNA METHYLATION IN COVID-19 DISEASE

1193

Scott A Bowler[1], Georgios Papoutsoglou[2], Aristides Karanikas[2], Ioannis Tsamardinos[2,3], Michael J Corley[1], Lishomwa C Ndhlovu[1]
[1]Department of Medicine, Division of Infectious Disease, Weill Cornell Medicine, New York, NY, USA,
[2]JADBio – Gnosis DA S.A, Heraklion, Greece, [3]Department of Computer Science, University of Crete, Heraklion, Greece

BACKGROUND

SARS-CoV-2 infection has resulted in over 219 million confirmed cases of COVID-19 with 4.5 million fatalities, highlighting the importance of elucidating mechanisms of severe disease. Here we utilized machine learning (ML) technologies to ***identify DNA methylation (DNAm) footprints of COVID-19 disease*** from publicly available data.

METHODS

- Genome-wide DNAm of SARS-CoV-2-infected and –uninfected individuals was assessed by Illumina Infinium MethylationEPIC profiling array from whole blood was available through publicly available datasets
- **GSE167202** (training cohort, Table 1) comprised of 460 participants trichotomized as uninfected healthy controls (HC), Mild COVID-19 (MCV) - ICU not required, or Severe COVID-19 (SVC) - ICU admittance or death.
- **GSE174818** (external validation, Table 2) consisting of 102 COVID-19 infected and 26 with non-COVID-19-related pneumonia was used to evaluate the disease specificity of our model.
- M values were obtained through beta value logit transformation
- The JADBio automated machine learning platform was leveraged for predictive modeling.

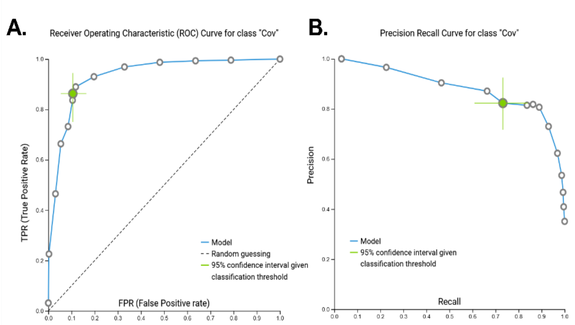
RESULTS

Table 1. Training Cohort (GSE167202) Demographics:

	Total Cohort	Healthy Control (HC)	Mild COVID-19 (MCV)	Severe COVID-19 (SCV)	p-value (HC v SCV)
N	460	296	118	46	
Age, years [IQR]	56 [39, 68]	59 [41, 70]	50 [37, 62]	50 [38, 62]	0.05
Gender, male (%)	274 (59.6%)	151 (51.0%)	63 (53.4%)	30 (65.2%)	0.07
Race, n (%)					0.001
African American	92 (20.0%)	50 (16.9%)	19 (16.1%)	6 (13.0%)	
Asian	18 (3.9%)	6 (2.0%)	8 (6.8%)	4 (8.7%)	
Caucasian	268 (58.3%)	184 (62.2%)	27 (22.9%)	19 (41.3%)	
Other	82 (17.8%)	56 (18.9%)	64 (54.2%)	17 (37.0%)	
Ethnicity, n (%)					
Hispanic or Latinx	157 (34.1%)	57 (19.3%)	68 (57.6%)	20 (43.5%)	<0.001
COVID Hospitalization, Days [IQR]	NA	NA	3 [0.5, 6]	18 [10, 31]	NA
Mortality, n (%)	21 (4.6%)	7 (2.4%)	2 (1.7%)	10 (21.7%)	<0.001

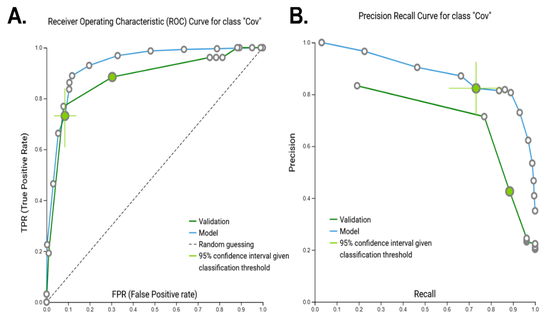
RESULTS

Figure 1. Model Performance (Training)



Receiver-operating characteristic (ROC) and Precision-Recall curves (PRC) for the performance of the artificial intelligence-based classification model to identify patients with severe COVID-19 disease in the training (GSE167202) cohort. AUC = area under the curve; "Cov" = COVID-19 disease classification. Colored circle represents the optimized classification threshold to predict COVID-19 disease, while bars extending from the circle represent the 95% confidence. A. AUC-ROC 0.933 [0.855, 0.970]. B. AUC-PRC 0.965 [0.932, 0.986]

Figure 2. Model Performance (External validation)



ROC and C for the performance of the artificial intelligence-based classification model to identify patients with COVID-19 disease from non-COVID-19-associated pneumonia (external validation [GSE174181]. A. AUC-ROC 0.901. B. AUC-PRC 0.748.

LASSO feature selection identified 4 unique methylation sites (summarized in **Table 3**) capable of carrying predictive weight for a 10-fold cross validated classification random forest trained on 100 trees with deviance splitting criterion (minimum leaf size=3). The average area under the curve (AUC) of receiver operator characteristics (ROC) of the model was 0.933 (95% confidence interval [0.885, 0.970]) while the AUC-Precision-recall curve (PRC) was 0.965 [0.932, PR0.986] **(Figure 1).** When applied to our external validation, this model produced an AUC-ROC of 0.901 with an AUC-PRC of 0.748 (**Figure 2**).

Table 2. External Validation Cohort (GSE174181) Demographics:

	Total Cohort	Non-COVID-19 (Non)	COVID-19 (Cov)	p-value (Cov vs Non)
n	128	26	102	
Age, years [IQR]	63 [51, 75]	65 [53, 75]	62 [50, 74]	0.496
Gender, male (%)	77 (60.2%)	13 (50.0%)	64 (62.7%)	0.238
Race, n (%)				0.005
African American	15 (11.7%)	4 (15.4%)	11 (10.8%)	
Asian	2 (2.6%)	0 (0%)	2 (2.0%)	
Caucasian	67 (52.3%)	21 (80.8%)	46 (45.1%)	
Hispanic or Latinx	22 (17.2%)	1 (3.8%)	21 (20.6%)	
Other	22 (17.2%)	0 *0%)	22 (21.6%)	
ICU admittance, n (%)	67 (52.4%)	16 (61.5%)	51 (50.0%)	0.295

Table 3. Genome associated CpG sites

	Location	Gene	Reported exposure
cg10778971	chr14:94577101	IFI27	Age, pregnancy-associated hypertension, cancer
cg07878065	chr18:2641871	-	Maternal dysglycemia, mixed connective tissue disease
cg03753191	chr13:43566902	EPSTI1	Age, primary Sjorgren's syndrome, rheumatoid arthritis
cg17114584	chr11:613792	IRF7	Age, primary Sjorgren's syndrome, smoking, multiple sclerosis

IFI27 – Interferon Alpha Inducible Protein 27
EPSTI1 – Epithelial Stromal Interaction 1
IRF7 – Interferon Regulatory Factor 7

SUMMARY

4 unique methylation sites carried predictive power for a random forest classification with an ***AUC-ROC of 0.933 and AUC-PRC of 0.965*** in HC vs SVC. When applied to our external validation, the model produced an ***AUC-ROC of 0.901 with an AUC-PRC of 0.748***

CONCLUSIONS

Our analyses revealed a 4 CpG signature with the predictive power to classify individuals with severe COVID-19 disease by leveraging JADBio AutoML platform. 3 of the genes associated with these CpG sites (IFI27, EPSI1, and IRF7) have been shown to be upregulated from scRNA-seq data. To the best of our knowledge, we are the first to report on cg07878065 in the context of SARS-CoV-2 infection. These results enhance our understand of the epigenetics mechanisms used by SARS-CoV-2 in disease pathogenesis and identify potential therapeutic targets.

ACKNOWLEDGMENTS