AAAS
Health Data Science
Volume 2021, Article ID 7574903, 9 pages
https://doi.org/10.34133/2021/7574903

Health Data Science
A SCIENCE PARTNER JOURNAL

Research Article

Machine Learning Highlights Downtrending of COVID-19 Patients with a Distinct Laboratory Profile

He S. Yang,[1,2] **Yu Hou**,[3] **Hao Zhang**,[3] **Amy Chadburn**,[1,2] **Lars F. Westblade**,[1,2,4] **Richard Fedeli**,[2] **Peter A. D. Steel**,[2,5] **Sabrina E. Racine-Brzostek**,[1,2] **Priya Velu**,[1,2] **Jorge L. Sepulveda**,[6] **Michael J. Satlin**,[7] **Melissa M. Cushing**,[1,2] **Rainu Kaushal**,[2,3] **Zhen Zhao**,[1,2] **and Fei Wang**[3]

[1]*Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA*
[2]*New York-Presbyterian Hospital/Weill Cornell Medical Campus, New York, NY, USA*
[3]*Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA*
[4]*Department of Medicine, Weill Cornell Medicine, New York, NY, USA*
[5]*Department of Emergency Medicine, Weill Cornell Medicine, New York, NY, USA*
[6]*Department of Pathology, School of Medicine and Health Sciences, George Washington University, Washington DC, USA*
[7]*Division of Infectious Disease, Department of Medicine, Weill Cornell Medicine, New York, NY, USA*

Correspondence should be addressed to He S. Yang; hey9012@med.cornell.edu, Zhen Zhao; zhz9010@med.cornell.edu, and Fei Wang; few2001@med.cornell.edu

Received 3 December 2020; Accepted 7 February 2021; Published 16 June 2021

Background. New York City (NYC) experienced an initial surge and gradual decline in the number of SARS-CoV-2-confirmed cases in 2020. A change in the pattern of laboratory test results in COVID-19 patients over this time has not been reported or correlated with patient outcome. *Methods*. We performed a retrospective study of routine laboratory and SARS-CoV-2 RT-PCR test results from 5,785 patients evaluated in a NYC hospital emergency department from March to June employing machine learning analysis. *Results*. A COVID-19 high-risk laboratory test result profile (COVID19-HRP), consisting of 21 routine blood tests, was identified to characterize the SARS-CoV-2 patients. Approximately half of the SARS-CoV-2 positive patients had the distinct COVID19-HRP that separated them from SARS-CoV-2 negative patients. SARS-CoV-2 patients with the COVID19-HRP had higher SARS-CoV-2 viral loads, determined by cycle threshold values from the RT-PCR, and poorer clinical outcome compared to other positive patients without the COVID12-HRP. Furthermore, the percentage of SARS-CoV-2 patients with the COVID19-HRP has significantly decreased from March/April to May/June. Notably, viral load in the SARS-CoV-2 patients declined, and their laboratory profile became less distinguishable from SARS-CoV-2 negative patients in the later phase. *Conclusions*. Our longitudinal analysis illustrates the temporal change of laboratory test result profile in SARS-CoV-2 patients and the COVID-19 evolvement in a US epicenter. This analysis could become an important tool in COVID-19 population disease severity tracking and prediction. In addition, this analysis may play an important role in prioritizing high-risk patients, assisting in patient triaging and optimizing the usage of resources.

1. Introduction

The coronavirus disease-2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1], has rapidly spread across the globe resulting in 110 million confirmed cases and 2.45 million total deaths as of February 20, 2021 [2]. The United States has more confirmed cases than any other country worldwide. New York, which was the initial epicenter of the COVID-19 pandemic and has reported the highest number of death in the US [3], has experienced a gradual decline in the number of cases in the months following the initial surge [4, 5]. It is unclear if the decline in total Emergency Department (ED) visits for COVID-19-like illnesses [6] and COVID-19-associated

hospitalizations [7] is related to changes in virus virulence, early preferential infection of more vulnerable populations, effectiveness of containment measures, or treatment changes. However, there have been only limited studies describing trends in objective clinical data in COVID-19 patients corresponding to these epidemiologic changes.

Currently in most hospital EDs, patients with symptoms suspicious for COVID-19 undergo a SARS-CoV-2 reverse transcription-polymerase chain reaction (RT-PCR) test and a panel of routine laboratory tests. While the pathophysiology of this new virus is still poorly understood, some of its effects on the human body are reflected in abnormal laboratory values. Several studies [8–10] have reported a number of abnormal routine laboratory test results in SARS-CoV-2-infected patients upon initial evaluation, including changes in the complete blood count (CBC), an increase in inflammatory markers and alterations in albumin and globulin levels. Whether the laboratory characteristics of SARS-CoV-2-infected patients have also shifted with the epidemiological changes over time, reflecting the evolution of COVID-19, remains unknown.

Machine learning algorithms have been successfully utilized in healthcare [11–13] and are powerful tools for predicting SARS-CoV-2 infection status [10, 14], disease progression, and mortality [15]. They are particularly useful in disentangling the hidden relationships among complex sets of variables. As routine laboratory test results provide objective and quantifiable characterization of the effects of the virus on the human body, our study is aimed at elucidating the trending of COVID-19 from a laboratory testing prospective. Using machine learning analysis, we identified a distinct panel of abnormal test results (COVID-19 high-risk laboratory test result profile; COVID19-HRP), which separate SARS-CoV-2 positive from SARS-CoV-2 negative patients. Our analysis visualized the temporal changes in the laboratory characteristics of SARS-CoV-2 positive patients from the initial outbreak in March and April to a postapex phase in May and June 2020 which may provide epidemiological insights to basic scientists and policy makers.

2. Methods

2.1. Patient Cohort and Data Collection. The test results analyzed in this study were from 5,785 patients evaluated in the ED of a New York academic hospital from March 11 to June 30, 2020 (Institute Review Board #20-03021671). SARS-CoV-2 RT-PCR results, routine laboratory testing results, patient demographic information (age, sex, and race, Table 1), and clinical outcome (hospital admission, ICU admission, mechanical intubation, and survival/death) were obtained from the laboratory information system (Cerner Millennium, Cerner Corporation, North Kansas City, Missouri, US). Since the turn-around time (TAT) of RT-PCR is up to 24 hours in our institution whereas the results of routine laboratory testing are usually available within 1-2 hours, laboratory testing results performed within a 48-hour window (±24 hours) of completion of each RT-PCR test were used in the data analysis. Exclusion criteria included patients <18 years old, patients who had indeterminate RT-PCR results (RT-PCR positive for the pan-Sarbecovirus target (E gene), yet negative for the SARS-CoV-2-specific target), and patients who did not have any laboratory test results within the time frame (inclusion/exclusion cascade, Figure 1). In total, our dataset included the routine laboratory test results from 1,309 SARS-CoV-2 RT-PCR positive and 3,658 RT-PCR negative patients (total 4,967 patients) who ranged in age from 18 to 104 years (median = 60 years). Violin plots of the age distribution in all patients as well as SARS-CoV-2 positive patients during the 4 study months are shown in Figure 2.

2.2. SARS-CoV-2 RT-PCR Testing. SARS-CoV-2 RT-PCR testing was performed using the RealStar SARS-CoV-2 RT-PCR Kit 1.0 reagent system (Altona, Hamburg, Germany) which targets on the S gene and E gene, the Cobas SARS-CoV-2 Assay (Roche Molecular Systems, Inc., Branchburg, NJ) which targets the ORF1ab and E genes, and the Xpert Xpress SARS-CoV-2 Assay (Cepheid, Inc., Sunnyvale, CA) which targets the N2 and E genes [16]. The ORF1ab and N2 genes are specific for SARS-Cov-2, while the E gene is a pan-Sarbecovirus marker. Based on the previous data [16], the diagnostic performance of both the Cobas 6800 and the Xpert Xpress SARS-CoV-2 assays are considered equivalent. SARS-CoV-2 RT-PCR cycle threshold (C_T) values of the SARS-CoV-2-specific target, which correlate inversely with the quantitative viral load [17], were obtained using the Cobas SARS-CoV-2 Assay and Xpert Xpress SARS-CoV-2 RT-PCR Assay, as the values for the SARS-CoV-2-specific gene were comparable between platforms [16]. C_T values from the RealStar SARS-CoV-2 RT-PCR assay were excluded from the analysis as the values are not directly comparable to the other two platforms.

2.3. Routine Laboratory Testing. Routine chemistry testing was performed on the Siemens ADVIA XPT and Centaur XP analyzers (Siemens Healthineers Global, Erlangen, Germany). Procalcitonin was performed on the Roche e411 analyzer (Roche Diagnostics, Indianapolis, IN). Blood gas analysis was performed on the GEM Premier 4000 analyzer (Instrumentation Laboratory, Bedford, MA). Routine hematology testing was performed on the UniCel DXH 800 analyzer (Beckman Coulter, Brea, CA). Coagulation tests were performed on the Instrumentation Laboratory ACLTM TOP CTS Coagulation System.

2.4. The Unified Manifold Approximation and Projection (UMAP) Analysis. Twenty-one laboratory tests were selected from a total of 685 tests that were ordered for all patients in the dataset based on the following criteria: (1) the test result was available for at least 70% of the patients within a 48-hour window around a specific SARS-CoV-2 RT-PCR test in each month and (2) the test result was significantly different (i.e., p value, p value after Bonferroni correction, or p value after demographics adjustment less than 0.05) in patients with a positive SARS-CoV-2 RT-PCR study compared to persons who had a negative result (Table 1). If one specific test was ordered multiple times within 48 hours, an

Table 1: Demographic information of the patient cohort and comparison of 21 laboratory tests in SARS-CoV-2 positive and negative patients.

Feature name	p value	p value (Bonferroni correction)	p value (demographics adjustment)	Total patients	Positive patients (n = 1,309)	Negative patients (n = 3,658)
Male n (%)	1.35e-15	-	-	2,415 (48.62%)	748 (30.97%)	1,667 (69.03%)
Female n (%)	-	-	-	2,552 (51.38%)	537 (21.04%)	2,015 (78.96%)
Age Mean (SD)	3.53e-17	-	-	58.33 (20.36)	62.61 (17.61)	56.83 (21.03)
Race: Black or African American n (%)	0.05 (White vs. Black)	-	-	530 (10.67%)	122 (23.02%)	408 (76.98%)
Race: Asian	-	-	-	240 (4.83%)	49 (20.42%)	191 (79.58%)
Race: Caucasian	-	-	-	1,691 (34.04%)	321 (18.98%)	1,370 (81.02%)
Race: other	-	-	-	2,506 (50.45%)	793 (31.64%)	1,713 (68.36%)
Anion gap Median (25%-75% quantile)	7.80e-27	2.11e-25	8.45e-16	9.0 (7.7, 11.0)	9.5 (8.5, 11.0)	9.0 (7.5, 10.5)
Albumin (g/dL)	1.31e-89	3.53e-88	1.94e-68	3.4 (2.9, 3.9)	3.0 (2.7, 3.5)	3.6 (3.0, 4.1)
Alkaline phosphatase (U/L)	1.37e-09	3.69e-08	2.11e-02	79.0 (63.0, 107.0)	75.0 (58.5, 102.0)	80.5 (65.0, 108.5)
Indirect bilirubin (mg/dL)	2.35e-39	6.33e-38	5.79e-12	0.4 (0.3, 0.5)	0.3 (0.2, 0.4)	0.4 (0.3, 0.6)
Calcium (mg/dL)	1.94e-156	5.25e-155	2.92e-123	9.0 (8.5, 9.5)	8.5 (8.1, 9.0)	9.15 (8.7, 9.6)
Chloride (mmol/L)	1.41e-32	3.80e-31	7.16e-17	103.5 (100.7, 106.0)	102.0 (99.0, 105.0)	104.0 (101.0, 106.0)
Globulin (g/dL)	1.75e-50	4.73e-49	3.45e-27	3.1 (2.8, 3.6)	3.4 (3.0, 3.7)	3.1 (2.7, 3.5)
Glucose (mg/dL)	2.52e-14	6.79e-13	7.73e-07	108.7 (95.5, 131.0)	113.0 (98.0, 141.4)	107.0 (95.0, 128.7)
Sodium (mmol/L)	1.44e-37	3.88e-36	1.55e-18	139.0 (137.0, 141.0)	138.0 (135.5, 140.3)	139.5 (137.5, 141.1)
Total protein (g/dL)	1.59e-16	4.29e-15	5.02e-11	6.6 (6.1, 7.1)	6.5 (6.0, 6.9)	6.7 (6.2, 7.2)
Basophil percentage	5.91e-60	1.60e-58	7.27e-31	0.45% (0.30%, 0.70%)	0.30% (0.20%, 0.50%)	0.50% (0.30%, 0.70%)
Hematocrit	1.07e-12	2.88e-11	5.33e-11	37.85% (33.49%, 41.42%)	38.6% (35.15%, 42.10%)	37.54% (32.86%, 41.20%)
Hemoglobin (g/dL)	1.79e-16	4.83e-15	7.50e-15	12.6 (11.0, 13.9)	12.9 (11.6, 14.2)	12.4 (10.8, 13.8)
White blood cell (WBC) (×10(3)/μL)	2.97e-28	8.02e-27	5.54e-03	7.90 (5.90, 10.60)	6.92 (5.10, 9.55)	8.10 (6.20, 10.9)
Lymphocyte count (×10(3)/μL)	2.43e-69	6.56e-68	3.17e-04	1.20 (0.78, 1.75)	0.90 (0.60, 1.25)	1.31 (0.86, 1.90)
Mean corpuscular volume (MCV, fL)	4.13e-07	1.12e-05	6.13e-07	89.7 (85.7, 93.5)	89.0 (85.1, 92.5)	90.0 (85.9, 93.9)
Monocyte count (×10(3)/μL)	3.34e-37	9.02e-36	8.96e-25	0.6 (0.4, 0.8)	0.5 (0.4, 0.7)	0.6 (0.5, 0.8)
Neutrophil count (×10(3)/μL)	1.37e-06	3.70e-05	1.37e-05	5.4 (3.7, 8.0)	5.1 (3.4, 7.5)	5.5 (3.8, 8.2)
Red blood cell count (×10(6)/μL)	4.50e-19	1.22e-17	2.73e-17	4.24 (3.72, 4.69)	4.41 (3.92, 4.79)	4.19 (3.66, 4.66)
Red blood cell distribution width (RDW-CV)	3.74e-07	1.01e-05	5.39e-12	14.3% (13.4%, 15.8%)	14.0% (13.4%, 15.2%)	14.3% (13.4%, 16.0%)
Magnesium (mg/dL)	1.43e-04	3.87e-03	5.80e-03	1.97 (1.80, 2.11)	2.00 (1.80, 2.20)	1.95 (1.80, 2.10)

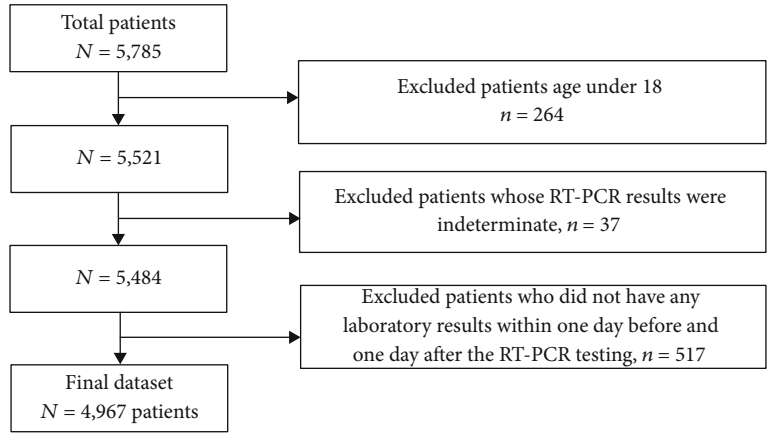
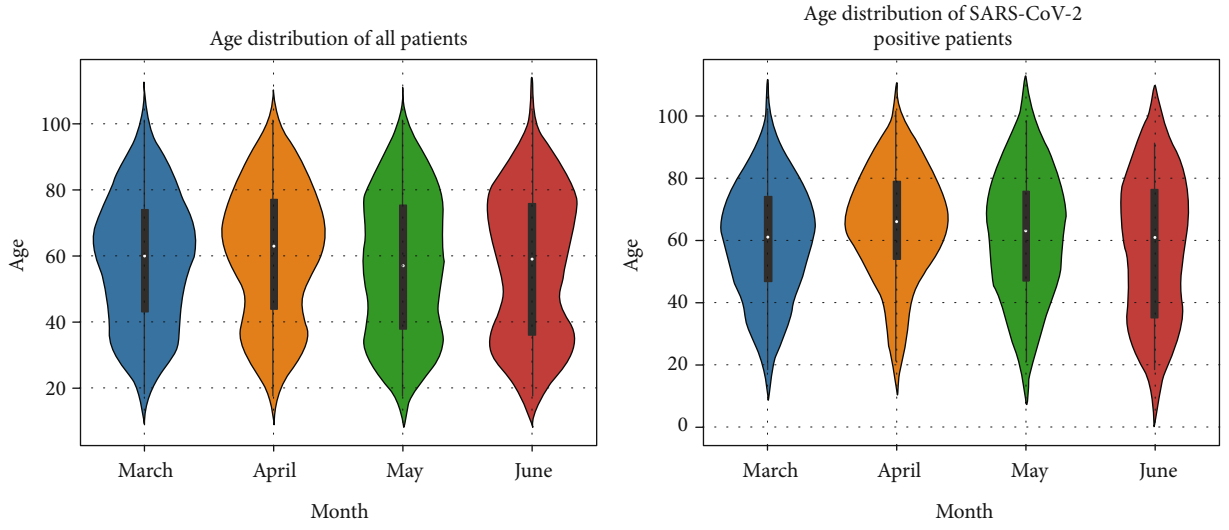


Figure 1: Inclusion/exclusion cascade of patients in the dataset.



	Mean (SD)	Median (25% -75% quantile)
March	58.66 (18.76)	60.0 (44, 73)
April	60.80 (19.69)	63.0 (45, 76)
May	56.93 (20.68)	57.0 (38.75, 74)
June	57.29 (21.30)	59.0 (37, 75)

(a)

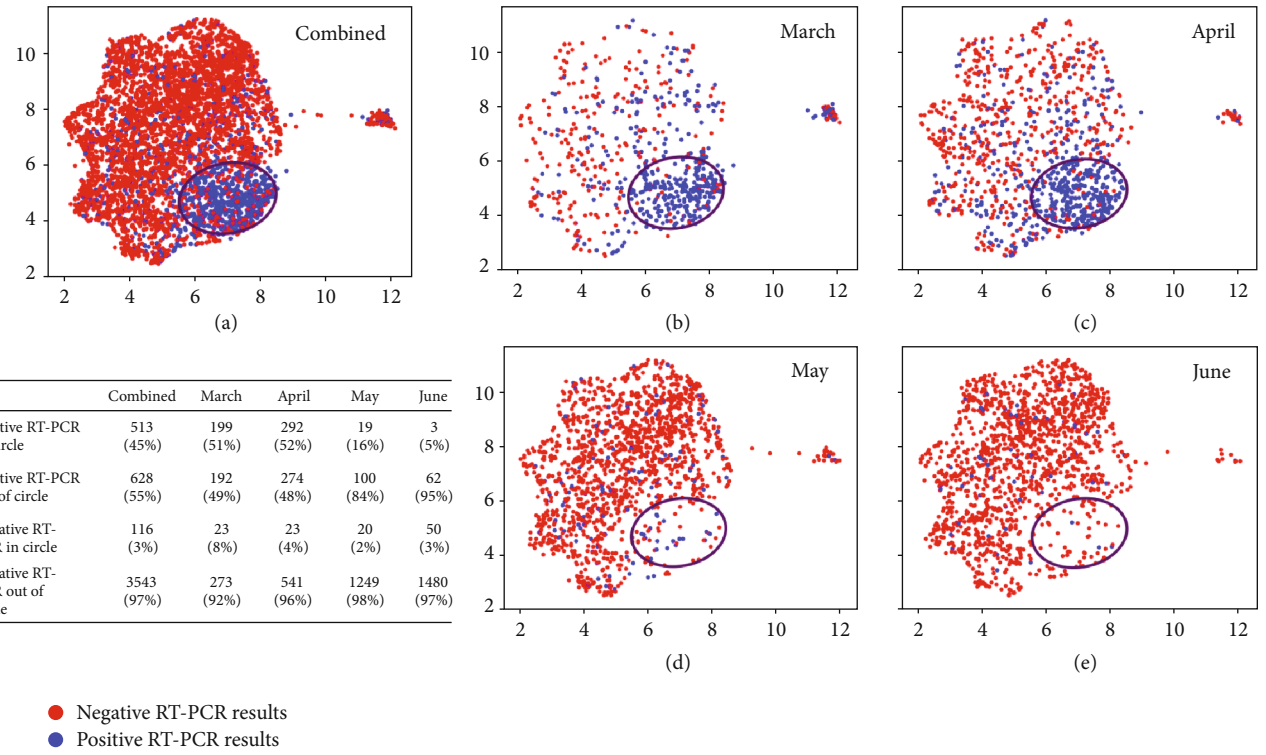
	Mean (SD)	Median (25% -75% quantile)
March	59.87 (17.29)	61.0 (48, 73)
April	65.11 (16.89)	66.0 (55, 78)
May	61.79 (18.76)	63.0 (48, 74.75)
June	58.19 (20.92)	61.0 (36, 75.25)

(b)

Figure 2: Distribution of age in total RT-PCR tested patients (a) and SARS-CoV-2 positive patients (b) in March, April, May, and June. Mean (SD) and median (25%-75% quantile) are shown under each figure.

average of the values was calculated and used for analysis. The missing value of a specific laboratory test in a feature vector was imputed by the median value of the available non-missing values of that dimension over all patients. Finally, a 21-dimensional vector was constructed to represent every SARS-CoV-2 RT-PCR testing result, which is a unique laboratory test result profile that characterizes each patient.

We then mapped the vectors of all RT-PCR tests onto a two-dimensional space using the UMAP approach [18], with the goal of visualizing the geometric distributions of the RT-PCR test profiles. UMAP is a dimensionality reduction technique aiming at projecting the data samples in a low-dimensional space such that the geometric sample relationships in the original high-dimensional feature space are preserved. Therefore, the geometric relationships among the sample vectors can be visually inspected. These profiles were first standardized with z-score scaling [19] before being incorporated into the UMAP algorithm to eliminate the value range discrepancies among different routine laboratory tests. Therefore, RT-PCR results with similar routine



	Combined	March	April	May	June
Positive RT-PCR in circle	513 (45%)	199 (51%)	292 (52%)	19 (16%)	3 (5%)
Positive RT-PCR out of circle	628 (55%)	192 (49%)	274 (48%)	100 (84%)	62 (95%)
Negative RT-PCR in circle	116 (3%)	23 (8%)	23 (4%)	20 (2%)	50 (3%)
Negative RT-PCR out of circle	3543 (97%)	273 (92%)	541 (96%)	1249 (98%)	1480 (97%)

Figure 3: Unified Manifold Approximation and Projection (UMAP) analysis of the laboratory profiles associated with the RT-PCR SARS-CoV-2-positive and SARS-CoV-2-negative testing results during March, April, May, and June combined (a), as well as separately in March (b), April (c), May (d), and June (e). Blue and red dots represent positive and negative RT-PCR results, respectively. The black circle depicts the high-density positive RT-PCR region. The singleton cluster on the right of the UMAP embedding includes 105 patients with 90% feature values missing in their profile vectors. Those missing values are imputed as the overall mean of each feature, which makes those profiles almost identical to each other. Since UMAP preserves the pairwise similarity during the mapping process, these vectors are mapped to a tiny crowd, which was excluded from our next analysis. Percentage of positive RT-PCR within and outside the circle and percentage of negative RT-PCR within and outside the circle are shown in the table, respectively.

laboratory profiles remain nearby in the embedding space whereas those with distinct laboratory profiles are located at a distance.

After all, RT-PCR profiles were projected onto the two-dimensional space, we used Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [20] to identify the high-density region of positive tests. DBSCAN is a density-based clustering algorithm that can automatically identify the high-density regions in the sample space without strong assumptions or the need for specifying the optimum number of clusters. Then we fitted a two-dimensional Gaussian distribution to define a circle in the two-dimensional embedding space. The mean vector and covariance matrix of this Gaussian distribution is [7.01, 4.76] and [(0.55 0.06), (0.06 0.41)], respectively. After having the Gaussian distribution, we plotted its contour lines for probability density function (pdf). Starting from the contour line with the largest pdf value (0.33), which is the mean point, we gradually expanded the contour line with a decrease of the pdf value in a step size of 0.01. In this expanding process, if we found that the number of negative tests was larger than that of positive ones, we would stop and regard this contour line as the circle.

2.5. Statistical Analysis. Comparison of the percentages of RT-PCR results within versus outside the circle in each month was performed by Fisher's exact test and post hoc analysis. Comparison of the C_T values and length of hospital stay within versus outside the circle was performed by *t*-test. Comparison of the percentage of SARS-CoV-2 positive patients with or without the circle for hospital admission from ED, percentage of patients required for care in the ICU and mechanical intubation were performed by the Fisher's exact test, where the *p* values were obtained after age adjustment. Statistical analysis was performed using Python version 3.7.

3. Results

A retrospective analysis of laboratory tests was performed in a final dataset of 1,309 SARS-CoV-2 RT-PCR-confirmed positive patients and 3,658 negative patients (Figure 1). A summary of the 21 laboratory tests used to construct the 21-dimensional vector representing the COVID19-HRP is shown in Table 1. Using the UMAP analysis, we then mapped the vectors of 5,588 RT-PCR tests onto a two-dimensional space. As shown in Figure 3, 45% ($n = 513$) of the overall SARS-CoV-2 RT-PCR-positive results clustered in the area within the black circle which depicts the high-density region of positive RT-PCR results. The patients who had positive RT-PCR results within the circle showed a

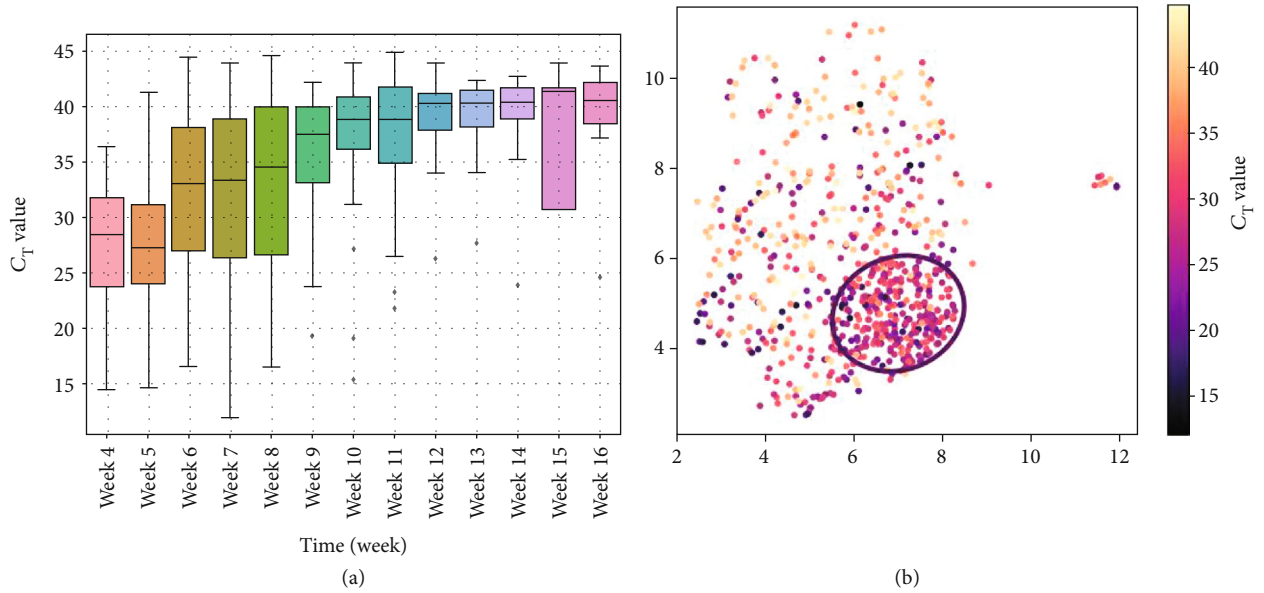


FIGURE 4: Trend of the SARS-CoV-2 RT-PCR cycle threshold (C_T) values for the SARS-CoV-2 specific target. (a) Box plot the C_T values in each week from April to June. (b) UMAP analysis of the C_T value associated with the SARS-CoV-2 RT-PCR results. The black circle is the same as in Figure 2. Color bar shows the SARS-CoV-2 RT-PCR C_T value from low (black) to high (yellow).

high-risk laboratory test result profile (COVID19-HRP) different from those individuals with negative RT-PCR results. In contrast, only 3% ($n = 116$) of SARS-CoV-2-negative RT-PCR results shared the COVID19-HRP and were within the circle. We further performed the UMAP analysis for each of the four months (March, April, May, and June) and observed a dramatic change over time: approximately half of the RT-PCR positive results in March (51%) and April (52%) clustered within the circle. When transitioning into May, while the total number of positive cases was declining, positive RT-PCR results associated with COVID19-HRP became significantly fewer, with only 16% of positive RT-PCR results in the circle ($p < 0.001$ compared to March or April, respectively). In June, the percentage of SARS-CoV-2 RT-PCR positive results in the circle was even less (5%, $p = 0.03$ compared to May) and relatively more positive RT-PCR were indistinguishably intermixed with the negative RT-PCR results based on the laboratory test result profile. However, it is important to note that more than 90% of the SARS-CoV-2 RT-PCR-negative results (97% overall, 92% in March, 96% in April, 98% in May, and 97% in June) fell outside the circle throughout the initial and subsequent months of the SARS-CoV-2 pandemic.

To characterize the COVID19-HRP, we investigated the distribution of each laboratory test corresponding to the positive and negative RT-PCR results within and outside the circle, respectively. Violin plots of representative laboratory tests (Supplemental Figure S1) showed, for example, that COVID-19 patients presenting in the ED, as part of the COVID19-HRP, had lower absolute lymphocyte and monocyte counts, lower percentage of basophils, hypocalcemia, and higher red blood cell counts as well as higher hemoglobin levels and hematocrits compared to the SARS-CoV-2-negative ED patients. While no single laboratory test can accurately discriminate SARS-CoV-2 infected from uninfected patients, the combination of 21 laboratory tests formed a distinct profile that characterized typical SARS-CoV-2-positive ED patients, separating them from the SARS-CoV-2-negative ED patients.

As shown in Figure 4, overall, the C_T values of SARS-CoV-2 RT-PCR results demonstrated an increasing trend (i.e., decreasing viral load) from April to June (C_T values in March were excluded from the analysis as they were generated from the Altona RealStar instrument with values that were not directly comparable with the other RT-PCR instruments [21]). The RT-PCR results within the circle had lower C_T values compared to those outside the circle (mean ± SD: 28.3 ± 5.0 vs. 32.4 ± 7.6, median: 28.7 vs. 33.0, $p < 0.001$). In other words, higher viral loads were seen in SARS-CoV-2-positive patients who had the COVID19-HRP compared to other positive patients who did not.

Chart reviews were performed to investigate the clinical outcome of each SARS-CoV-2 positive patient, including whether they were discharged from the ED or admitted to an inpatient ward, whether they required care in the ICU, whether they developed respiratory failure and were intubated, and whether they died or survived COVID-19. Twenty-one patients who were transferred to other hospitals were excluded due to unknown outcomes. Overall, SARS-CoV-2 positive patients with the COVID19-HRP in our dataset had a higher incidence of hospital admission (95.7% vs. 78.4%, $p < 0.001$), ICU admission (27.2% vs. 15.2%, $p < 0.001$), and intubation (24.7% vs. 11.5%, $p < 0.001$) than SARS-CoV-2 patients without the COVID19-HRP, where the p values were obtained after age adjustment. For the patients who had been admitted, the length of in-hospital stay was significantly longer in SARS-CoV-2 patients with the COVID19-HRP than the other positive patients without the COVID19-HRP (mean ± SD: 16.6 ± 22.1 vs. 12.7 ± 21.0, median 8 vs. 5 days, $p < 0.001$).

We further investigated the patients who had negative RT-PCR results but had laboratory testing results that mapped within the circle ($n = 116$). Among them, 48 patients presented to the ED with COVID-19-like symptoms such as fever, cough, dyspnea, and/or malaise, and 3 were reported to have close contacts with persons who tested positive for SARS-CoV-2. Nine patients (7.8% of the 116 patients) were diagnosed with COVID-19 within two days upon repeated RT-PCR testing (majority of patients tested negative did not have a repeated testing) and four other patients (2.5%) tested positive for COVID-19 antibodies one to two months after their ED visit. Therefore, the combination of specific laboratory testing results may identify some SARS-CoV-2-infected patients with a false negative RT-PCR result. Three patients were diagnosed with another respiratory virus infection such as influenza A or human rhinovirus/enterovirus.

4. Discussion

In this study, using machine learning analysis, we show that approximately half of the SARS-CoV-2-positive ED patients had a distinct profile of routine laboratory test results that clearly separate them from the SARS-CoV-2-negative patients. Notably, the SARS-CoV-2 patients with the COVID19-HRP had an overall higher viral load and poorer clinical outcome compared to the other positive patients without the COVID19-HRP. The identification of COVID-19 distinct laboratory profile could be used to prioritize high-risk patients, assisting in ED patient triaging and optimizing the usage of resources in areas where RT-PCR testing is not accessible due to financial or supply constraints. Furthermore, our temporal analysis illustrates the substantial decrease in the percentage of patients with the COVID19-HRP in May and June 2020, after the initial surge of COVID-19 in March and April 2020, in NYC. The observed trend in the laboratory result profile provides insight to the epidemiologic and biologic evolution of the disease, which could play an important role in COVID-19 population disease severity tracking and prediction and may assist in directing public health policies as COVID-19 spreads to new geographic areas or as a resurgence occurs in previously affected areas.

Existing research has shown that the SARS-CoV-2 viral load correlates with severity of COVID-19 presentation [22] and is independently associated with an increased risk of intubation and/or in-hospital mortality [23–25]. Here, we demonstrate that SARS-CoV-2 viral load also correlates with a panel of laboratory test result abnormalities (COVID19-HRP). Patients who have a higher viral load and a COVID19-HRP at presentation may have a higher risk of adverse outcomes. Thus, our analysis provides a means of identifying patients with more severe physiologic disturbance and poorer outcome. Analysis of the laboratory profile at ED presentation provides complementary information, which, because of the rapid turn-around time (usually within a couple of hours) for routine laboratory test results, offers an opportunity for rapid triaging and more timely intensive monitoring of high-risk patients. In addition, this analysis may also suggest which patients are unlikely to be SARS-CoV-2 positive, as overall 97% of SARS-CoV-2-negative patients were outside the circle (did not have the COVID19-HRP). As such, this analysis could be deployed clinically as an application integrated into the electronic medical record (EMR) system and visually show if the dot corresponding to an individual patient is within or outside the circle as soon as the patient's laboratory test results are available. In areas where SARS-CoV-2 RT-PCR is not accessible onsite, this analysis may provide a timely clue to prioritize high-risk patients.

Laboratory tests provide an objective and quantifiable means to characterize the evolution of COVID-19. In addition to an overall decrease in the number of positive cases, our study depicts a declining trend in the viral load of SARS-CoV-2 patients as well as a decreasing percentage of patients showing the COVID19-HRP from April to June 2020. In our hospital, RT-PCR tests in March and April were primarily offered to symptomatic patients due to a limited testing capacity. Testing was expanded to more patients, both symptomatic and asymptomatic, in May and June when supplies, equipment, and testing personnel were available. While more widely available testing in May and June may contribute to the decrease in the percentage of severe patients, it is unclear whether there are other contributing factors such as changes in virus virulence, modifications of population behavior by adhering to public health directives such as wearing masks, increased patient awareness of the disease with physician visits sooner after symptom onset (presumably associated with lower viral loads), and a decrease in the number of most vulnerable patients as they have already been infected. Our analysis, based upon a patient population in NYC during the peak of COVID-19, provides to researchers, physicians, and public health authorities an insightful method to better understand the evolution of this disease from a laboratory testing perspective. In addition, our model based on laboratory test results reflecting the physiologic effects of the virus on patients may improve our understanding of the pathobiology of the SARS-CoV-2, and thus, aid in devising guidance for treatment, tracking, and prevention of COVID-19. Another indication of our study is the need for model updating and retraining for risk prediction of COVID-19 patients in different time periods of the pandemic.

Our study has a limitation that the analysis of patient data was performed at a single large metropolitan medical center. Therefore, the role of the COVID19-HRP in discriminating between SARS-CoV-2-negative and SARS-CoV-2-positive patients should be tested on a larger scale at other medical centers in areas with varying degrees of COVID-19 prevalence. In addition, due to the reagent and consumable supply shortages, our institute stopped testing for influenza and other respiratory viruses from March to September 2020. Therefore, we were not able to analyze the laboratory test result profiles in patients who had other respiratory viruses. Differentiating SARS-CoV-2 from other respiratory virus infections will be one of our future studies.

5. Conclusions

Using machine learning analysis, we have identified a typical laboratory test result profile for SARS-CoV-2 positive

patients, which correlates with higher viral load and poorer clinical outcome. Overall, 97% of the SARS-CoV-2 negative patients did not have the COVID19-HRP. This analysis could serve as an important tool to prioritize high-risk patients and optimize the usage of resource. Furthermore, this analysis illustrates the downtrending in the proportion of SARS-CoV-2 patients with the COVID19-HRP from the initial surge of COVID-19 to a later postapex phase in NYC, the initial epicenter of the pandemic in the US. Our findings have shed new light on the evolution and pathobiology of COVID-19.

Abbreviations

COVID-19:	Coronavirus disease-2019
COVID19-HRP:	High-risk COVID-19 laboratory test result profile
C_T:	Cycle threshold
DBSCAN:	Density-based spatial clustering of applications with noise
ED:	Emergency department
ICU:	Intensive care unit
MCV:	Mean corpuscular volume
RDW-CV:	Red blood cell distribution width
RT-PCR:	Reverse transcription-polymerase chain reaction
SARS-CoV-2:	Severe acute respiratory syndrome coronavirus 2
TAT:	Turn-around time
UMAP:	Unified manifold approximation and projection
WBC:	White blood cell.

Conflicts of Interest

None of the authors has a conflict of interest in this project.

Authors' Contributions

HSY is responsible for conceptualization, investigation, data collection and analysis, writing the original draft, and editing of the manuscript. YH and HZ are responsible for data analysis, visualization, and editing the manuscript. AC and LFW are responsible for conceptualization and editing the manuscript. RF is responsible for organizing the dataset and performing data analysis. SRB is responsible for editing the manuscript. PV is responsible for providing RT-PCR data. JLS is responsible for providing C_T values and editing the manuscript. MMC, RK, and ZZ are responsible for reviewing and editing the manuscript. FW is responsible for conceptualization, supervision of the project, investigation, data analysis, and editing the manuscript.

Acknowledgments

We want to thank Hanna Rennert and Arryn R. Craney for their effort on RT-PCR method development and medical technologists who performed the laboratory testing. The work of YH, HZ, and FW was partially supported by NSF 1750326 and 2027970 and ONR N00014-18-1-2585.

Supplementary Materials

Supplemental Figure S1: distribution of representative laboratory tests of positive RT-PCR within the TPR, negative RT-PCR within the TPR, positive RT-PCR outside the TPR, and negative RT-PCR outside the TPR. *(Supplementary Materials)*

References

[1] N. Zhu, D. Zhang, W. Wang et al., "A novel coronavirus from patients with pneumonia in China, 2019," *The New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020.

[2] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533-534, 2020.

[3] P. Goyal, J. J. Choi, L. C. Pinheiro et al., "Clinical characteristics of covid-19 in New York city," *The New England Journal of Medicine*, vol. 382, no. 24, pp. 2372–2374, 2020.

[4] New York City Health, "*COVID-19: data*," 2021, https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

[5] New York City HealthOctober 2020, https://www1.Nyc.Gov/site/doh/covid/covid-19-data.Page.

[6] C. N. Thompson, J. Baumgartner, C. Pichardo et al., "Outbreak - New York City, February 29-June 1, 2020," *Morbidity and Mortality Weekly Report*, vol. 69, no. 46, pp. 1725–1729, 2020.

[7] Centers for Disease Control and Prevention, *Laboratory-confirmed covid-19-associated hospitalization*, 2020, https://giscdcgov/grasp/COVIDNet/COVID19_3html.

[8] W. J. Wiersinga, A. Rhodes, A. C. Cheng, S. J. Peacock, and H. C. Prescott, "Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19): a review," *JAMA*, vol. 25, pp. 782–793, 2020.

[9] W. J. Guan, Z. Y. Ni, Y. Hu et al., "Clinical characteristics of coronavirus disease 2019 in China," *The New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[10] H. S. Yang, Y. Hou, L. V. Vasovic et al., "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning," *Clinical chemistry*, vol. 66, no. 11, pp. 1396–1404, 2020.

[11] N. Tomašev, X. Glorot, J. W. Rae et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[12] R. D. Ganetzky and S. R. Master, "Machine learning for the biochemical genetics laboratory," *Clinical Chemistry*, vol. 66, no. 9, pp. 1134-1135, 2020.

[13] E. L. Gill and S. R. Master, "Hidden in plain sight: machine learning in acute kidney injury," *Clinical Chemistry*, vol. 66, no. 4, pp. 509–511, 2020.

[14] X. Y. Mei, H. C. Lee, K. Y. Diao et al., "Artificial intelligence-enabled rapid diagnosis of patients with covid-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[15] L. Yan, H. T. Zhang, J. Goncalves et al., "An interpretable mortality prediction model for COVID-19 patients," *Nature machine intelligence*, vol. 2, no. 5, pp. 283–288, 2020.

[16] D. A. Green, J. Zucker, L. F. Westblade et al., "Clinical performance of SARS-CoV-2 molecular testing," *Journal of Clinical Microbiology*, vol. 58, no. 8, article e00995, 2020.

[17] M. G. Becker, T. Taylor, S. Kiazyk, D. R. Cabiles, A. F. Meyers, and P. A. Sandstrom, "Recommendations for sample pooling on the cepheid genexpert® system using the cepheid xpert® xpress SARS-CoV-2 assay," *Plos one*, vol. 15, no. 11, p. e0241959, 2020.

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018, https://arxiv.org/abs/180203426.

[19] C. R. Wylie, *Advanced engineering mathematics University of Utah, USA, McCraw-Hill Book Company Inc*, Tosho Printing Co Ltd, Tokyo, Japan, 1960, Card Number: 59-13221.

[20] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) AAAI Press*, pp. 226–231, Portland, Oregon, USA, 1996.

[21] P. Velu, A. Craney, P. Ruggiero et al., "Rapid implementation of SARS-CoV-2 emergency use authorization RT-PCR testing and experience at an academic medical institute," *The Journal of molecular diagnostics*, vol. 23, 2020.

[22] S. Zheng, J. Fan, F. Yu et al., "Viral load dynamics and disease severity in patients infected with sars-cov-2 in Zhejiang province, China, January-March 2020: retrospective cohort study," *BMJ*, vol. 369, p. m1443, 2020.

[23] R. Magleby, L. F. Westblade, A. Trzebucki et al., "Impact of Severe Acute Respiratory Syndrome Coronavirus 2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019," *Clinical infectious diseases*, 2020.

[24] E. Pujadas, F. Chaudhry, R. McBride et al., "SARS-CoV-2 viral load predicts covid-19 mortality," *The Lancet Respiratory Medicine*, vol. 8, no. 9, article e70, 2020.

[25] L. F. Westblade, G. Brar, L. C. Pinheiro et al., "SARS-CoV-2 viral load predicts mortality in patients with and without cancer who are hospitalized with covid-19," *Cancer Cell*, vol. 38, no. 5, pp. 661–671.e2, 2020.