

Machine learning analysis highlights the down-trending of the proportion of COVID-19 patients with a distinct laboratory result profile

He S. Yang, PhD^{1,2}, Yu Hou, PhD³, Hao Zhang, PhD³, Amy Chadburn, MD^{1,2}, Lars F. Westblade, PhD^{1,2,4}, Richard Fedeli², Peter A.D. Steel, MD^{2,5}, Sabrina E. Racine-Brzostek, MD^{1,2}, Priya Velu, MD^{1,2}, Jorge L. Sepulveda, MD⁶, Michael J. Satlin, MD⁷, Melissa M. Cushing, MD^{1,2}, Rainu Kaushal, PhD^{2,3}, Zhen Zhao, PhD^{1,2}, Fei Wang, PhD^{3*}

¹Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

²New York-Presbyterian Hospital/Weill Cornell Medical Campus, New York, NY, US

³Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA

⁴Department of Medicine, Weill Cornell Medicine, New York, NY, US.

⁵Department of Emergency Medicine, Weill Cornell Medicine, New York, NY, USA

⁶Department of Pathology and Laboratory Informatics, School of Medicine and Health Sciences, George Washington University, Washington DC, US.

⁷Division of Infectious Disease, Department of Medicine, Weill Cornell Medicine, New York, NY, US

Corresponding authors:

Fei Wang, PhD

Division of Health Informatics

Department of Healthcare Policy and Research

Weill Cornell Medicine

425 E. 61th Street, New York City, NY 10065

Email: few2001@med.cornell.edu

Phone: 646-962-9405

He S. Yang, PhD

Assistant Professor

Department of Pathology and Laboratory Medicine

Weill Cornell Medicine

525 East 68th Street. New York, NY 10065

Email: hey9012@med.cornell.edu

Phone: 212-746-6292

Zhen Zhao, PhD

Associate Professor

Department of Pathology and Laboratory Medicine

Weill Cornell Medicine

525 East 68th Street. New York, NY 10065

Email: zhz9010@med.cornell.edu

Phone: 212-746-2682.

Key words: Coronavirus Disease 2019 (COVID-19), machine learning, Unified Manifold Approximation and Project (UMAP), routine laboratory tests, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Abbreviation: COVID-19: Coronavirus Disease-2019; COVID19-TLP: COVID-19 typical laboratory test result profile; C_T : cycle threshold; ED: emergency department; ICU: intensive care unit; MCV: mean corpuscular volume; NYPH/WCMC: New York Presbyterian Hospital/Weill Cornell Medical Center; 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 Project; WBC: white blood cell.

Running title: Down-trending of COVID-19 patients with a typical lab profile

Abstract:

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 COVID19-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 study visualized the down-trending of the proportion of SARS-CoV-2 patients with the distinct COVID19-HRP. 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.

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 40.8 million confirmed cases and 1.1 million total deaths as of October 21, 2020 (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 U.S. (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 applications for predicting SARS-CoV-2 infection status (10, 14), disease progression and mortality (15). They are particularly useful in identifying

hidden relationships based on 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 aimed to elucidate 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 typical laboratory test result profile; COVID19-TLP), which separate SARS-CoV-2 positive from SARS-CoV-2 negative patients and visualized the temporal changes in the COVID19-TLP of SARS-CoV-2 positive patients from the initial outbreak in March and April to a post-apex phase in May and June 2020.

Methods

Ethics Statement:

This study was approved by the Weill Cornell Medicine Institutional Review Board (#20-03021671).

Patient cohort and data collection:

The test results analyzed in this study were from 5,785 patients evaluated in the ED of New York Presbyterian Hospital/Weill Cornell Medical Center (NYPH/WCMC) from March 11 to June 30, 2020. 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, 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 (\pm 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 = 60.0 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 **Supplemental Figure 1**.

SARS-CoV-2 RT-PCR testing

SARS-CoV-2 RT-PCR testing was performed at NYPH/WCMC 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 upon 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.

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.

The 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 48 hours prior to or after 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 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. 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. The UMAP analysis allows the geometric topology among the high dimensional vectors to remain in the low dimensional space so that the geometric relationships among the sample vectors can be visually inspected. Therefore, RT-PCR results with similar routine 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. 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.

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 posthoc 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.

Role of funders:

The funders have no role in study design, data collection, data analysis, interpretation or writing of the manuscript.

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-TLP 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 2**, 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 distinct laboratory test result profile (COVID19-TLP) 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-TLP 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-TLP 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-TLP, 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 2**) show, for example that COVID-19 patients presenting in the ED, as part of the COVID19-TLP, had lower absolute lymphocyte, monocyte and basophil counts, 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 3**, 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 \pm 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-TLP 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-TLP 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-TLP, where the p-values were obtained after age adjustment. For the patients who had been admitted, the length of stay in hospital was significant longer

in SARS-CoV-2 patients with the TLP than the other positive patients without the TLP (mean \pm 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-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.

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-TLP had an overall higher viral load and poorer clinical outcome compared to the other positive patients without the COVID19-TLP. 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-TLP 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 “second wave” 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)(24, 25). Here, we demonstrate that SARS-CoV-2 viral load also correlates with a panel of laboratory test result abnormalities (COVID19-TLP). Patients who have a higher viral load and a COVID19-TLP 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-TLP). 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-TLP 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), 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 guidances for treatment, tracking and prevention of COVID-19.

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-TLP in discriminating between SARS-CoV-2 negative and positive patients should be tested on a larger scale at other medical centers in areas with varying degrees of COVID-19 prevalence.

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-TLP. This analysis could serve as an important tool to prioritize high-risk patients and optimize the usage of resource. Furthermore, this analysis illustrates the down-trending in the proportion of SARS-CoV-2 patients with the COVID19-TLP from the initial surge of COVID-19 to a later post-apex 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.

Declaration of interests

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

Acknowledgement

We want to thank Hanna Rennert and Arryn R. Craney for their effort on RT-PCR method development, and medical technologists at NYPH/WCMC who performed the laboratory testing.

Author contribution:

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

Funding/Support:

The work is partially supported by National Science Foundation under grant number 1750326 and 2027970, and Office of Naval Research under grant number N00014-18-1-2585

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	Positive (n=1309)	Negative (n=3658)
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 (7.67, 11.0)	9.5 (8.5, 11.0)	9.0 (7.5, 10.5)
Albumin	1.31E-89	3.53E-88	1.94E-68	3.4 (2.88, 3.9)	3.0 (2.65, 3.45)	3.6 (3.0, 4.05)
Alkaline phosphatase	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	2.35E-39	6.33E-38	5.79E-12	0.4 (0.25, 0.53)	0.3 (0.2, 0.4)	0.4 (0.3, 0.6)
Calcium	1.94E-156	5.25E-155	2.92E-123	9.0 (8.5, 9.5)	8.5 (8.12, 8.95)	9.15 (8.7, 9.6)
Chloride	1.41E-32	3.80E-31	7.16E-17	103.5 (100.67, 106.0)	102.0 (99.0, 105.0)	104.0 (101.0, 106.0)

Globulin	1.75E-50	4.73E-49	3.45E-27	3.13 (2.77, 3.55)	3.37 (3.0, 3.7)	3.05 (2.7, 3.5)
Glucose	2.52E-14	6.79E-13	7.73E-07	108.67 (95.5, 131.0)	113.0 (98.0, 141.38)	107.0 (95.0, 128.67)
Sodium	1.44E-37	3.88E-36	1.55E-18	139.0 (137.0, 141.0)	138.0 (135.5, 140.33)	139.5 (137.5, 141.09)
Total Protein	1.59E-16	4.29E-15	5.02E-11	6.6 (6.1, 7.1)	6.47 (6.0, 6.92)	6.7 (6.15, 7.2)
Basophil percentage	5.91E-60	1.60E-58	7.27E-31	0.45 (0.3, 0.7)	0.3 (0.2, 0.5)	0.5 (0.3, 0.7)
Hematocrit	1.07E-12	2.88E-11	5.33E-11	37.85 (33.49, 41.42)	38.6 (35.15, 42.1)	37.54 (32.86, 41.2)
Hemoglobin	1.79E-16	4.83E-15	7.50E-15	12.55 (11.0, 13.85)	12.9 (11.6, 14.15)	12.4 (10.8, 13.77)
White blood cell (WBC)	2.97E-28	8.02E-27	5.54E-03	7.9 (5.9, 10.6)	6.92 (5.1, 9.55)	8.1 (6.2, 10.9)
Lymphocyte count	2.43E-69	6.56E-68	3.17E-04	1.2 (0.78, 1.75)	0.9 (0.6, 1.25)	1.31 (0.86, 1.9)
Mean corpuscular volume (MCV)	4.13E-07	1.12E-05	6.13E-07	89.7 (85.7, 93.5)	89.02 (85.1, 92.45)	89.95 (85.91, 93.9)
Monocyte count	3.34E-37	9.02E-36	8.96E-25	0.6 (0.4, 0.8)	0.5 (0.35, 0.7)	0.6 (0.45, 0.8)
Neutrophil count	1.37E-06	3.70E-05	1.37E-05	5.42 (3.7, 8.0)	5.1 (3.4, 7.5)	5.5 (3.8, 8.2)
Red blood cell count	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	14.25 (13.4, 15.8)	14.03 (13.37, 15.2)	14.3 (13.4, 16.0)
Magnesium	1.43E-04	3.87E-03	5.80E-03	1.97 (1.8, 2.11)	2.0 (1.8, 2.2)	1.95 (1.8, 2.1)

References:

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in china, 2019. *N Engl J Med* 2020;382:727-33.
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect Dis* 2020;20:533-4.
3. Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, et al. Clinical characteristics of covid-19 in new york city. *N Engl J Med* 2020;382:2372-4.
4. Health NYSDo. Covid-19: Data. Available at <https://www1.Nyc.Gov/site/doh/covid/covid-19-data.Page>. 2020.
5. New York City Health. <https://www1.Nyc.Gov/site/doh/covid/covid-19-data.Page>. Accessed 10/13/2020. .
6. Centers for Disease Control and Prevention. Emergency department visits percentage of visits for covid-19-like illness (cli) or influenza-like illness (ili). Available at <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/06052020/covid-like-illness.html> 2020.
7. Centers for Disease Control and Prevention. Laboratory-confirmed covid-19-associated hospitalization. Available at https://gis.cdc.gov/grasp/COVIDNet/COVID19_3.html 2020.
8. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19): A review. *JAMA* 2020;25:782-93.
9. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in china. *N Engl J Med* 2020;382:1708-20.
10. Yang HS, Hou Y, Vasovic LV, Steel P, Chadburn A, Racine-Brzostek SE, et al. Routine laboratory blood tests predict sars-cov-2 infection using machine learning. *Clin Chem* 2020;doi: 10.1093/clinchem/hvaa200.
11. Tomasev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019;572:116-9.
12. Ganetzky RD, Master SR. Machine learning for the biochemical genetics laboratory. *Clin Chem* 2020;66:1134-5.
13. Gill EL, Master SR. Hidden in plain sight: Machine learning in acute kidney injury. *Clin Chem* 2020;66:509-11.
14. Mei XY, Lee HC, Diao KY, Huang MQ, Lin B, Liu CY, et al. Artificial intelligence-enabled rapid diagnosis of patients with covid-19. *Nature Medicine* 2020;26:1224-8.
15. Yan L ZH, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for covid-19 patients. *Nat Mach Intell* 2020;2:283-8.
16. Green DA, Zucker J, Westblade LF, Whittier S, Rennert H, Velu P, et al. Clinical performance of sars-cov-2 molecular testing. *J Clin Microbiol* 2020;58(8):e00995-20. doi: 10.1128/JCM.-20.
17. Becker M.G. TT, Kiazzyk S., Cabiles D.R., Meyers A.F.A., Sandstrom P.A. . Recommendations for sample pooling on the cepheid genexpert® system using the cepheid xpert® xpress sars-cov-2 assay. *BioRxiv* <https://doi.org/10.1101/20200514097287> 2020.

18. McInnes L, HJ, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv: 180203426 2018.
19. C.R. W. Advanced engineering mathematics University of Utah, USA, McCraw-Hill Book Company Inc, Tosho Printing Co Ltd, Tokyo, Japan, Card Number: 59-13221 1960.
20. Ester M, KHP, Sander J, Xu X. . A density-based algorithm for discovering clusters in large spatial databases with noise. . Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) AAAI Press 1996:226-31. .
21. Velu P CA, Ruggiero P, Siple J, Cong L, Hissong E, Loda M, Westblade LF, Cushing M, Rennert H. Rapid implementation of sars-cov-2 emergency use authorization rt-pcr testing and experience at an academic medical institute. Medrxiv 2020.
22. Zheng S, Fan J, Yu F, Feng B, Lou B, Zou Q, 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 2020;369:m1443.
23. Magleby R, Westblade LF, Trzebucki A, Simon MS, Rajan M, Park J, et al. Impact of sars-cov-2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019. Clin Infect Dis 2020:doi: 10.1093/cid/ciaa851.
24. Pujadas E, Chaudhry F, McBride R, Richter F, Zhao S, Wajnberg A, et al. Sars-cov-2 viral load predicts covid-19 mortality. Lancet Respir Med 2020;8:e70. doi: 10.1016/S2213-600(20)30354-4. .
25. Westblade LF BGB, Pinheiro LC, Paidoussis D, Rajan M, Martin P, Martin P, et al. Sars-cov-2 viral load predicts mortality in patients with and without cancer who are hospitalized with covid-19. Cancer Cell 2020:10.1016/j.ccell.2020.09.007

Figure legends:

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

Figure 1

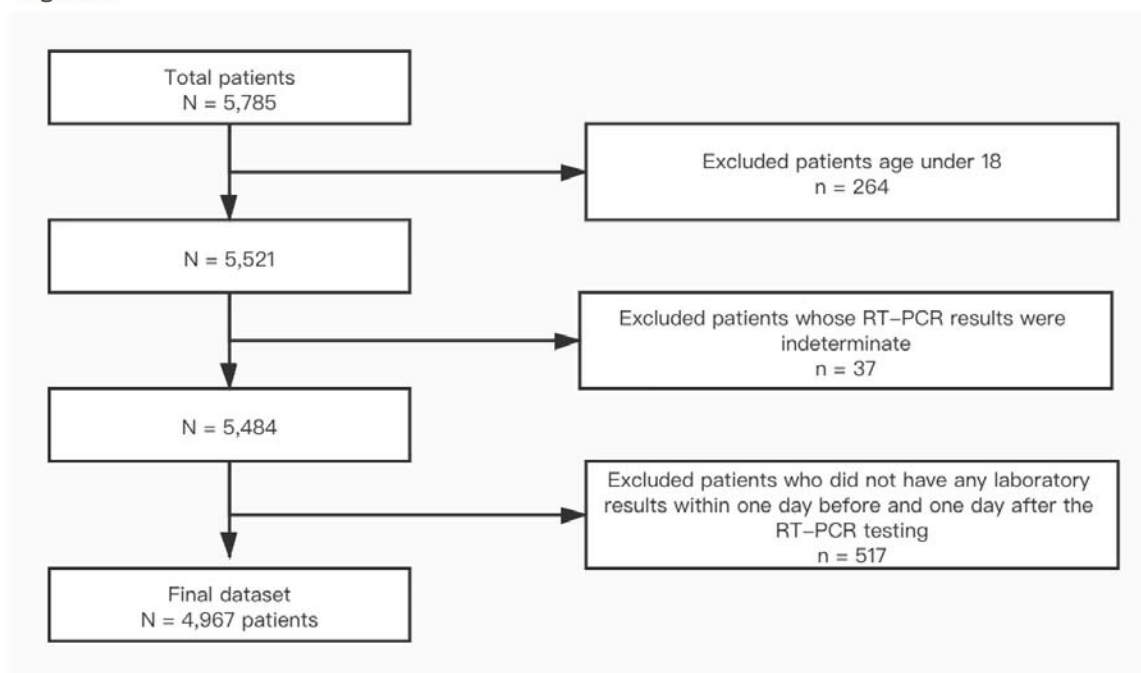


Figure 2. Unified Manifold Approximation and Projection (UMAP) analysis of the laboratory profiles associated with the SARS-CoV-2 RT-PCR positive and negative testing results during March, April, May and June combined (a), as well as separately in March (b), April (c), May (d) and June (e). Blue and 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 embeddings 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.

Figure 2

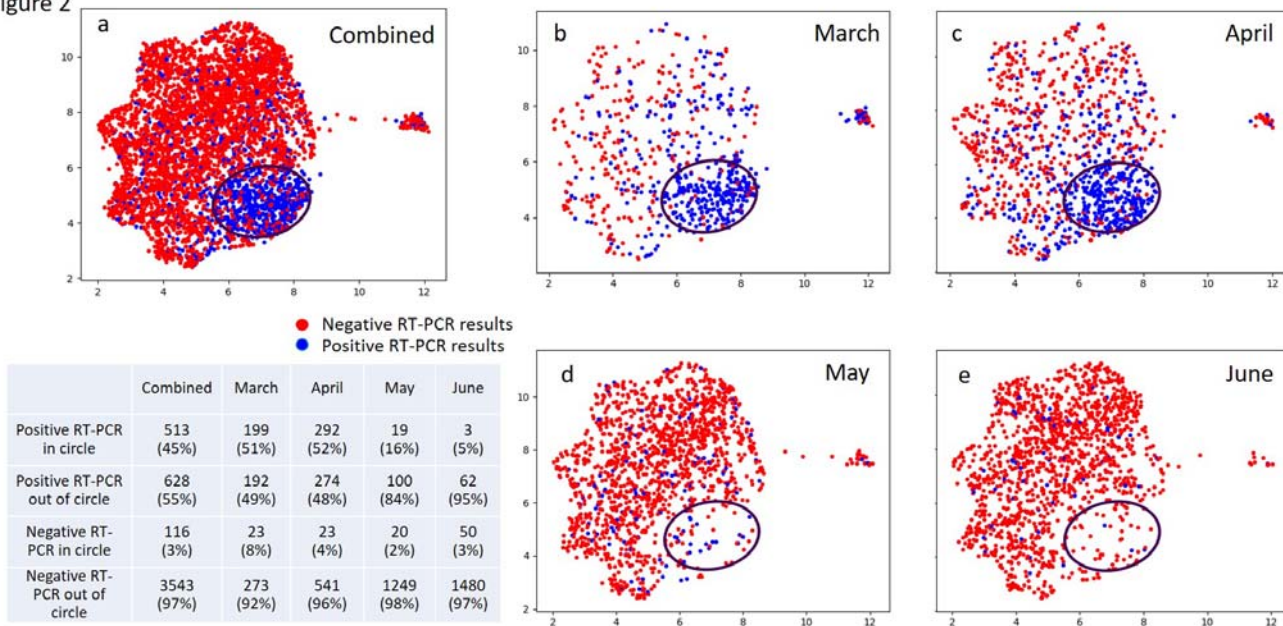
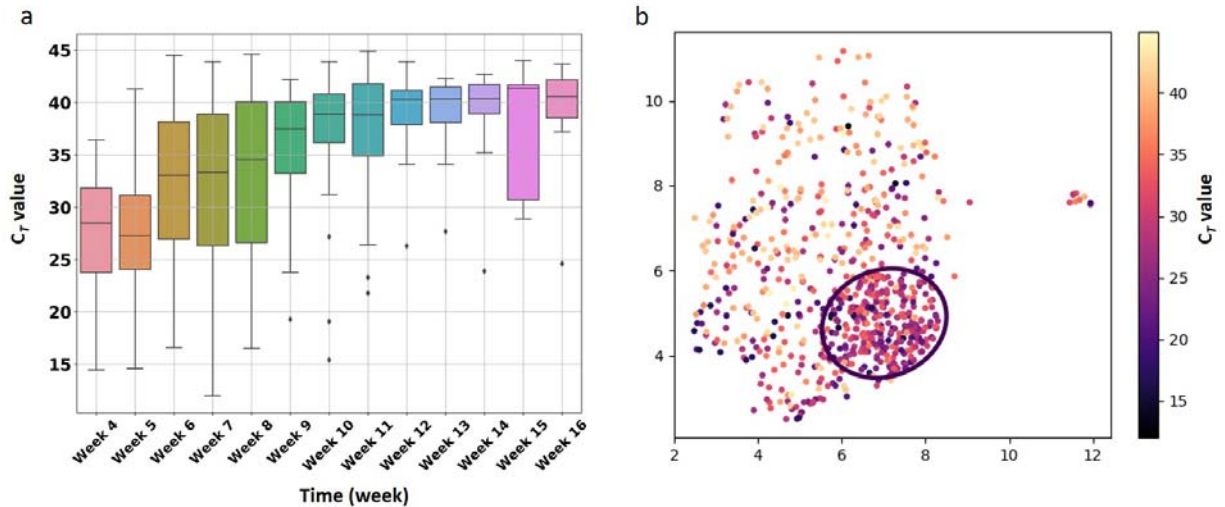
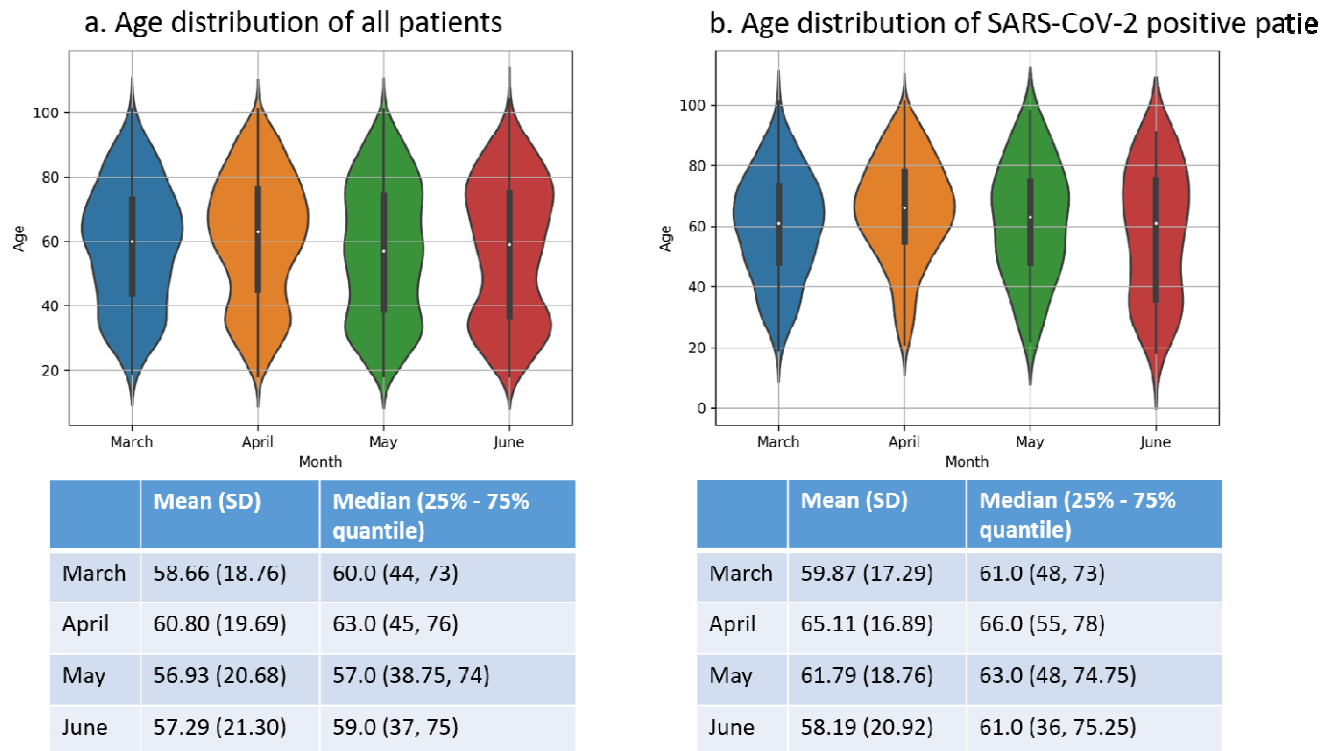


Figure 3. Trend of the SARS-CoV-2 RT-PCR cycle threshold (C_T) values for the SARS-CoV-2 specific target. (a) Box plot the C_T values in each week from April to June. (b) UMAP analysis of the C_T value associated with the SARS-CoV-2 RT-PCR results. The black circle is the same as in Figure 2. Color bar shows the SARS-CoV-2 RT-PCR C_T value from low (black) to high (yellow).

Figure 3



Supplemental Figure 1. 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.



Supplemental Figure 2. 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.

