A Machine Learning Model Incorporating Laboratory Blood Tests Discriminates Between SARS-CoV-2 and Influenza Infections at Emergency Department Visit

Junyi Liu¹, Lars F. Westblade^{2,3,4}, Amy Chadburn^{2,3}, Richard Fidelity^{2,3}, Arryn Craney^{2,3}, Sophie Rand²,
Melissa Cushing^{2,3}, Zhen Zhao^{2,3}, Jingjing Meng^{1*}, He S. Yang^{2,3*}

¹Department of Computer Science and Engineering, University of Buffalo, Buffalo, NY, USA

²Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

³New York-Presbyterian Hospital/Weill Cornell Medical College, New York, NY, USA

⁴Department of Medicine, Division of Infectious Diseases, Weill Cornell Medicine, New York, NY, USA

*Joint corresponding authors

Correspondence to:

He S. Yang, PhD

Department of Pathology and Laboratory Medicine

Weill Cornell Medicine

525 E. 68th Street, Suite F 707, New York, NY 10065

Email: hey9012@med.cornell.edu

Phone: 212-746-6292

20 Jingjing Meng, PhD

21 Department of Computer Science and Engineering

22 University of Buffalo

23 304 Davis Hall, Buffalo, NY

24 Email: Jmeng2@buffalo.edu

25 Phone: 716-645-0566

26

27

28 **Key words:** SARS-CoV-2, COVID-19, influenza, machine learning, random forest, routine laboratory tests

29 **Abbreviation:** AGAP: anion gap, COVID-19: corona virus disease-2019; SARS-CoV-2: severe acute

30 respiratory syndrome coronavirus 2; TAT: turn-around time; ED: emergency department; RT-PCR: real-

31 time reverse transcription polymerase chain reaction; WBC: white blood cells; RBC: red blood cells;

32 RDW-CV: Red blood cell distribution width; BUN: Blood urea nitrogen; MCV: Mean corpuscular volume;

33 AUC: Area under the receiver operating characteristic curve; NYPH/WCM: New York Presbyterian

34 Hospital/Weill Cornell Medicine; NYPH/LMH: New York Presbyterian Hospital/Lower Manhattan

35 Hospital.

36 **Running title:** Model discriminates COVID-19 and influenza

37 Word count: 2417

38

39

Abstract

Introduction: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and influenza virus are contagious respiratory pathogens with similar symptoms but require different treatment and management strategies. This study investigated whether laboratory blood tests can discriminate between SARS-CoV-2 and influenza infections at emergency department (ED) presentation.

Methods: 723 influenza A/B positive (2018/1/1 to 2020/3/15) and 1,281 SARS-CoV-2 positive (2020/3/11 to 2020/6/30) ED patients were retrospectively analyzed. Laboratory test results completed within 48 hours prior to reporting of virus RT-PCR results, as well as patient demographics were included to train and validate a random forest (RF) model. The dataset was randomly divided into training (2/3) and testing (1/3) sets with the same SARS-CoV-2/influenza ratio. The Shapley Additive Explanations technique was employed to visualize the impact of each laboratory test on the differentiation.

Results: The RF model incorporating results from 15 laboratory tests and demographic characteristics discriminated SARS-CoV-2 and influenza infections, with an area under the ROC curve value 0.90 in the independent testing set. The overall agreement with the RT-PCR results was 83% (95% CI: 80-86%). The test with the greatest impact on the differentiation was serum total calcium level. Further, the model achieved an AUC of 0.82 in a new dataset including 519 SARS-CoV-2 ED patients (2020/12/1 to 2021/2/28) and the previous 723 influenza positive patients. Serum calcium level remained the most impactful feature on the differentiation.

Conclusion: We identified characteristic laboratory test profiles differentiating SARS-CoV-2 and influenza infections, which may be useful for the preparedness of overlapping COVID-19 resurgence and future seasonal influenza.

Introduction:

Both the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiology agent of coronavirus disease 2019 (COVID-19), and influenza virus are contagious respiratory pathogens, which cause as a wide range of illnesses from asymptomatic or mild through to severe disease and death.¹ However, the fraction of severe infection cases, as well as the mortality rate, is higher in COVID-19 than in influenza.² Transmissibility, estimated by the basic reproductive rate (R_0), is also higher in SARS-CoV-2 ($R_0 \sim 2.5$) compared to influenza ($R_0 \sim 1.7$).³ The SARS-CoV-2 and influenza virus infections have similar symptoms, such as cough, sore throat, fever, fatigue, and myalgias, making clinical differentiation at hospital presentation challenging without the aid of laboratory tests. Currently, the detection of these pathogens relies on viral specific real-time reverse-transcription polymerase chain reaction (RT-PCR) testing of nasopharyngeal swabs or other upper respiratory track specimens.⁴ However, while the turn-around time (TAT) of RT-PCR testing is usually within a day, it can be substantially longer due to reagent shortages⁵ as experienced during the COVID-19 pandemic, or lack of on-site testing, particularly in many of the smaller and more rural hospitals. Since the treatment of SARS-CoV-2 and influenza is different, delayed identification of infected patients may result in inappropriate patient management and increased risk of infection for healthcare personnel. Therefore, rapid discrimination of SARS-CoV-2 and influenza infections, and identification of high-risk patients is vital for individual patient care and for controlling disease transmission.

Routine laboratory tests provide objective and quantifiable characterization of the effects of the virus on the human body⁶. Routine test results are generally available within 1-2 hours and are accessible prior to patient discharge from the emergency department (ED). Previous studies^{7, 8}, including our own publications⁹, have demonstrated that machine learning models incorporating routine laboratory blood tests and demographics can differentiate and predict SARS-CoV-2 infected and non-infected patients. Machine learning algorithms have the capability of revealing complicated pattern or trend behind high-

dimensional laboratory data that are challenging for human eyes to visualize. So far, limited studies have investigated the differences in laboratory blood test results between SARS-CoV-2 and influenza infections^{3, 10}. While no single laboratory test strongly associates with SARS-CoV-2 or influenza infection, we investigated whether a combination of multiple routine laboratory tests can discriminate between these two viral infections at an early stage of disease, and aimed to find out the most impactful laboratory features that contribute to the discrimination. Furthermore, COVID-19 has been evolving since the initial outbreak of the disease¹¹. We are interested in understanding whether the impactful laboratory features remain unchanged in more recent COVID-19 positive patients during the 2020-2021 winter wave.

Methods

Patient cohorts

We reviewed retrospectively laboratory and demographic data of 723 influenza A/B positive patients from January 1, 2018 to March 15, 2020, and 1,281 SARS-CoV-2 positive patients from March 11, 2020 to June 30, 2020, evaluated in the ED of a New York City academic hospital. Routine influenza RT-PCR testing was suspended from March to September 2020 to prioritize resources for SARS-CoV-2 testing. The SARS-CoV-2 cohort included infected patients from the initial outbreak as well as from the post-apex phase. Influenza or SARS-CoV-2 RT-PCR results, routine laboratory testing results within 2 days prior to the completion of RT-PCR testing, and patient demographic information were obtained from the laboratory information system (Cerner Millennium, Cerner Corporation). Exclusion criteria included patients < 18 years of age, and patients who lacked laboratory test results within the designated time window (within 2 days prior to RT-PCR results release). In total, the final dataset included results from 1,237 SARS-CoV-2 RT-PCR positive patients and 513 influenza positive (393 type A, and 119 type B, 1 A/B) ED patients (**Figure 1**). This study was approved by the Institutional Review Board (IRB) of Weill Cornell Medicine and deemed IRB exempt by the University of Buffalo.

From December 1, 2020 to February 28, 2021, there were a total of 559 SARS-CoV-2 RT-PCR positive patients and only 1 influenza RT-PCR positive patient (this patient was also positive for SARS-CoV-2) who were evaluated in the ED of our hospital. After applying the exclusion criteria as state above, 519 SARS-CoV-2 positive patients were included in the new dataset. Since there were only 1 positive influenza case during the same time, we had to compare the new SARS-CoV-2 dataset to the previous influenza dataset from 2018/1/1 to 2020/3/11.

Model construction

A total of 57 chemistry, hematology, and coagulation tests commonly ordered in these patients were examined in the SARS-CoV-2 and influenza cohorts. Of them, 15 laboratory tests were selected to be included in the model based on two criteria: 1) a result available for at least 70% of the patients within 48 hours before a specific SARS-CoV-2 RT-PCR test and 2) showing a significant difference (P-value after Bonferroni correction less than 0.05) between influenza and SARS-CoV-2 positive patients. After the feature selection, a 22-dimensional vector (15 laboratory tests, one age, one gender, five race variables (African American, Asian, Caucasian, others and unknown) was constructed to represent every SARS-CoV-2 or influenza RT-PCR test result (**Figure 2**). If one laboratory 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 mean value of the available non-missing values of that dimension over all patients. Subsequently, a random forest classifier model was developed incorporating the results of 15 selected laboratory tests and patient age, gender, and race, using the Python scikit-learn package 0.23.2. The whole data set was randomly split into a training set (2/3 of cases) and a testing set (1/3 cases) with the same ratio of SARS-CoV-2/influenza cases as the ratio for the overall cases. The hyperparameters of random forest model was trained using 5-fold cross validation, where all RT-PCR tests in the training set were randomly partitioned into 5 equal buckets with the same SARS-CoV-2/influenza ratio. The random forest model was then trained using selected hyperparameters with all RT-PCR tests in the training set. The performance of the model was evaluated in the independent testing set using area under the receiver operating characteristic curve (ROC), sensitivity, specificity, and agreement with RT-PCR. The Shapley Additive Explanations (SHAP) technique was employed to visualize the impact of each laboratory test on the differentiation.

Results:

Patients with SARS-CoV-2 infection were significantly older (median: 64 years, IQR: 51-75) than those with influenza (median: 48 years, IQR: 32-63, $p < 0.001$). The percentage of male patients was also significantly higher in SARS-CoV-2 cohort (58.36%) than in the influenza group (41.33%, $p < 0.001$). In addition, a significantly lower percentage of African American was present in the SARS-CoV-2 group than the influenza cohort ($p = 0.02$, **Table 1**).

The laboratory results completed within 48 hours prior to a positive SARS-CoV-2 or influenza RT-PCR test result were included in the analysis. A total of 15 laboratory test results with missing rate $< 30\%$ were found to be significantly different between the SARS-CoV-2 and influenza infected patients at their ED presentation. The performance of the random forest model incorporating these 15 laboratory tests and patient age, sex and race achieved an area under the ROC curve (AUC) value of 0.90 (95% CI: 0.88-0.93) in the independent testing set. The overall agreement with the RT-PCR, determined at the Youden's index point, was 83% (95% CI: 80-86%), with 87% (95% CI: 84-90%) agreement with SARS-CoV-2 RT-PCR results and 73% (95% CI: 66-80%) agreement with the influenza RT-PCR results. The random forest model generated a probability score, from 0 to 1, to differentiate influenza vs. SARS-CoV-2 infection (cutoff 0.58 at the Youden' index point).

In order to visualize and interpret the impact of each laboratory test to the overall differentiation, a Shapley Additive Explanations (SHAP) technique (Version 4.10.0) was employed which assigns each feature a value of importance (the SHAP value) for the differentiation of SARS-CoV-2 or influenza infections. A summary plot is shown in **Figure 2**, in which the impact of each laboratory test and demographic feature was ordered from high to low. Notably, serum calcium level had the greatest impact on the differentiation of these two infections. The SARS-CoV-2 patients demonstrated significantly lower serum calcium concentrations than the influenza positive patients. In addition, the peripheral monocyte and basophil counts were significantly lower whereas the lymphocyte and white

blood cell counts were significantly higher in the SARS-CoV-2 patients compared to the influenza patients. The SARS-CoV-2 patients also demonstrated hypochloremia, hyperglycemia, higher BUN/creatinine ratio and higher anion gap compared to the influenza patients.

Since the COVID-19 has been evolving since the initial outbreak, we further investigated whether our random forest model is still discriminative between SARS-COV-2 and influenza in the 2020-2021 winter wave and whether the top laboratory tests are still impactful on the differentiation. From December 1, 2020 to February 28, 2021, there were a total of 559 SARS-CoV-2 RT-PCR positive patients evaluated in the ED of our hospital, and among them, 519 were included in our dataset after the exclusion criteria was applied (Demographic information in **Table 1**). However, there was only one positive influenza case identified in the ED during this time period, and this patient was also positive for SARS-CoV-2 RT-PCR testing. Therefore, we test our model on the new SARS-CoV-2 dataset with the previous 723 influenza A/B positive patients from January 1, 2018 to March 15, 2020. The random forest model with the same parameters achieved an AUC of 0.82 with an overall agreement of 74% (95% CI: 71-77%) using the same cutoff value. Agreement with SARS-CoV-2 and influenza RT-PCR results was 75% and 73%, respectively.

Interestingly, serum calcium level remains the most impactful laboratory feature on the differentiation of SARS-CoV-2 and influenza infections (**Figure 3**). Compared to the influenza positive patients, the more recent SARS-CoV-2 positive patients still exhibited hypocalcemia, lower level of monocyte counts, and higher levels of lymphocyte, WBC, and neutrophil counts as well as higher anion gap and BUN/creatinine ratio.

Discussion:

This retrospective analysis illustrates that a combination of 15 laboratory tests, along with patient demographics, can help distinguish between SARS-CoV-2 and influenza infections at an early stage of disease. While SARS-CoV-2 and influenza infections show similar symptoms², each has a characteristic laboratory test result profile that could be used for early disease differentiation and prediction. Although the level of seasonal influenza was extremely low in the U.S. and other countries in the northern hemisphere during the past winter^{12, 13}, our findings pave the way for future research about the distinct pathophysiological mechanisms underlying SARS-CoV-2 and other respiratory virus infections, and may be useful for the preparedness of overlapping COVID-19 resurgence and future seasonal influenza. Understanding the characteristic laboratory profiles associated with the two pathogens may assist early and rapid identification of high-risk patients and optimize resource use during periods of high infection rates for both SARS-CoV-2 and influenza, especially in areas where hospitals or clinics do not have onsite RT-PCR testing for respiratory viruses.

Machine learning models can perform analysis of massive quantities of data and extract hidden patterns that that would be challenging for human eyes to identify. A machine learning classification model should make clinical and biologic sense. Here we report that in our machine learning classification model, the serum total calcium level, among all routine laboratory tests, has the highest impact on the differentiation of SARS-CoV-2 from influenza infections. Hypocalcemia has been previously reported as a prevalent biochemical abnormality in SARS-CoV-2 infected patients with a marked negative influence on disease severity, biochemical inflammation and thrombotic markers¹⁴, and has been proposed as an independent risk factor for hospitalization due to COVID-19¹⁵. While the exact underlying biologic mechanism for hypocalcemia in SARS-CoV-2 infection is still unclear, it may be closely associated with viral-associated multi-organ damage and increase in inflammatory cytokines¹⁶. Studies better delineating the mechanism and impact of viral infections on calcium metabolism may suggest

therapeutic modalities in the future. In addition, SARS-CoV-2 patients exhibited lower monocyte and basophil counts, higher white blood cell, neutrophil and lymphocyte counts, as well as hypochloremia, hyperglycemia, higher BUN/creatinine ratio and higher anion gap, compared to influenza positive patients. The substantial differences in laboratory result profile between SARS-CoV-2 and influenza patients shed new lights on the reactions of human body to the two types of viral infections, which may contribute to the different disease manifestations and outcomes.

The COVID-19 winter wave in New York was not as severe as the devastating first wave that hit the city in March and April 2020. The number of hospitalized patients as well as the mortality rates were lower than those of the first wave¹⁷. More patients showed mild or asymptomatic disease, which may partly explain why it was more difficult to differentiate SARS-CoV-2 and influenza infections from laboratory test results and why the performance of our model with the same parameters has dropped in the new dataset. The performance of the model can be improved with a continuous learning process that involves model updating and parameter optimizing using the new data. Interestingly, despite the evolvement of COVID-19, the laboratory features that have greatest impact on the differentiation between the two viral infections have been stable in the longitudinal datasets.

A study limitation is that our model's performance has not been validated in a dataset including concurrent SARS-CoV-2 and influenza positive patients as Influenza RT-PCR testing was suspended from March to September 2020 to prioritize resources for SARS-CoV-2 testing. We attempted to collect new data from November 2020 to February 2021, however, there was only one influenza positive case during this time in our hospital ED. This observation was consistent with the extremely low level of seasonal influenza in North America¹². Despite a lack of direct comparison, the characteristic profile of SARS-CoV-2 in comparison to influenza infection is still valid and has the potential to impact patient care. The performance of our model could be further improved when it is trained with more concurrent influenza and SARS-CoV-2 patient data.

Conclusion:

The proposed random forest model incorporating age, gender, race and 15 routine laboratory tests can discriminate between SARS-CoV-2 and influenza infection during an initial patient ED visit. We identified 15 routine laboratory blood tests, which assist in separating of SARS-CoV-2 from influenza infection, with serum calcium level being the most impactful feature. These characteristic laboratory test result profiles associated with SARS-CoV-2 and influenza infections mirror the biologic effects of these viruses on patients and provides clinically an opportunity for early identification of high-risk patients. Our analysis demonstrates the utility of machine learning as an emerging technique to support the diagnosis of infectious diseases.

Author contribution:

JL for performing data analysis and manuscript editing; LFW for providing influenza data and manuscript editing; AC for manuscript editing; RF for organizing laboratory test results for COVID-19 and influenza patients; AC for manuscript editing; SR for providing COVID-19 data; MC and ZZ for manuscript editing; JM for supervision of data analysis and manuscript editing; HSY for conceptualization, project supervision, data analysis, and writing and editing of the manuscript.

Conflict of Interest Disclosure:

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

Funding:

There is no funding source for this study.

Table 1. Demographic of SARS-CoV-2 and influenza patient cohorts

Demographics:	SARS-CoV-2 positive patients from 2020/3/11 to 2020/6/30 (n = 1,237)	Influenza positive patients from 2018/1/1 to 2020/3/15 (n = 513)	SARS-CoV-2 positive patients from 2020/12/1 to 2021/2/28 (n = 519)
Age, years; median (IQR)	64 (51-75)	48 (32-63) ^{*#}	61 (45-73)
Gender, male, n (%)	722 (58.36%)	212 (41.33%) ^{*#}	253 (48.75%)
Race:			
Asian	52 (4.20%)	32 (6.24%)	33 (6.36%)
African American	117 (9.46%)	66 (12.87%) [*]	71 (13.68%)
White	324 (26.19%)	153 (29.82%) [#]	21 (40.66%)
Other	213 (17.22%)	128 (24.95%) ^{*#}	93 (17.92%)
Unknown	531 (42.93%)	134 (26.12%) ^{*#}	111 (21.39%)

^{*} p value < 0.05 between SARS-CoV-2 positive patients from 2020/3/11 to 2020/6/30 and influenza positive patients from 2018/1/1 to 2020/3/15.

[#] p value < 0.05 between SARS-CoV-2 positive patients from 2020/12/1 to 2021/2/28 and influenza positive patients from 2018/1/1 to 2020/3/15.

Figure 1

Figure 2

279 **Figure 3.**

280

Figure legends:

Figure 1: Inclusion/exclusion cascade of SARS-CoV-2 and influenza patient cohorts in the dataset.

Figure 2: Impact of each laboratory test and demographic feature to the overall differentiation of SARS-

CoV-2 and influenza infections in the random forest model. Dataset includes SARS-CoV-2 positive

patients from 2020/3/15 to 2020/6/30, and influenza positive patients from 2018/1/1 to 2020/3/11.

Laboratory tests are organized on the y-axis according to their impact from high to low. Individual values

of each test for each patient are colored according to their relative values. Higher and lower values are

shown in red and blue, respectively. Absolute SHAP values of each test are shown on the x-axis. Positive

SHAP value to the right indicates prediction of SARS-CoV-2 whereas negative SHAP value to the left

indicates prediction of influenza.

Figure 3: Impact of each laboratory test and demographic feature to the overall differentiation in the

new SARS-CoV-2 dataset from 2020/12/1 to 2021/2/28 and the previous influenza dataset from

2018/1/1 to 2020/3/11.

References

1. World Health Organization (WHO). Coronavirus disease (COVID-19): Similarities and differences with influenza. URL: <https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-similarities-and-differences-with-influenza>. 2020.
2. Petersen E, Koopmans M, Go U, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect Dis*. 2020;20:e238-e244.
3. Manzanares-Meza LD, Medina-Contreras O. SARS-CoV-2 and influenza: a comparative overview and treatment implications. *Bol Med Hosp Infant Mex*. 2020;77:262-273.
4. Centers for Disease Control and Prevention (CDC). Overview of testing for SARS-CoV-2 (COVID-19); URL: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html#TestingInfection>. Accessed June 8, 2021.
5. American Society of Microbiology. Shortages of COVID-19 and other testing supplies identified by ASM's data collection tool. URL: <https://asm.org/Press-Releases/2020/October/Shortages-of-COVID-19-and-Other-Testing-Supplies-I>. Accessed on May 26, 2021.
6. Yang HS, Hou Y, Zhang H et al. Machine learning highlights down-trending of COVID-19 patients with a distinct laboratory profile. *Health Data Science*. 2021; Article ID 7574903. Doi: 10.34133/7574903.
7. Bayat V, Phelps S, Ryono R, et al. A SARS-CoV-2 Prediction Model from Standard Laboratory Tests. *Clin Infect Dis*. 2020; ciaa1175. doi: 10.1093/cid/ciaa1175.
8. Joshi RP, Pejaver V, Hammarlund NE, et al. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. *J Clin Virol*. 2020;129:104502.
9. Yang HS, Hou Y, Vasovic LV, et al. Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning. *Clin Chem*. 2020;66:1396-1404.

10. Song X, Delaney M, Shah RK, Campos JM, Wessel DL, DeBiasi RL. Comparison of Clinical Features of COVID-19 vs Seasonal Influenza A and B in US Children. *JAMA Netw Open*. 2020;3:e2020495.
11. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. *Exp Mol Med*. 2021;53:537-547.
12. Olsen SJ, Azziz-Baumgartner E, Budd AP, et al. Decreased Influenza Activity During the COVID-19 Pandemic - United States, Australia, Chile, and South Africa, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69:1305-1309.
13. Jones N. How COVID-19 is changing the cold and flu season. *Nature News Feature* 2020. URL: <https://www.nature.com/articles/d41586-020-03519-3>. Accessed June 8, 2020.
14. Di Filippo L, Doga M, Frara S, Giustina A. Hypocalcemia in COVID-19: Prevalence, clinical significance and therapeutic implications. *Rev Endocr Metab Disord*. 2021; doi: 10.1007/s11154-021-09655-z.
15. Di Filippo L, Formenti AM, Rovere-Querini P, et al. Hypocalcemia is highly prevalent and predicts hospitalization in patients with COVID-19. *Endocrine*. 2020;68:475-478.
16. Zhou X, Chen D, Wang L, et al. Low serum calcium: a new, important indicator of COVID-19 patients from mild/moderate to severe/critical. *Biosci Rep*. 2020; 40(12): BSR20202690.
17. Lewis C DJ. Why New York's last COVID surge was far less deadly than its first. *Gothamist*. URL: <https://gothamist.com/news/why-new-yorks-last-covid-surge-was-far-less-deadly-than-its-first>. Last accessed: June 28, 2021. 2021.