Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study

Timothy B Plante, Aaron Blau, Adrian N Berg, Aaron S Weinberg, Ik E Jun, Victor F Tapson, Tanya S Kanigan, Artur Adib

Submitted to: Journal of Medical Internet Research
on: September 01, 2020

Table of Contents

Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study

Timothy B Plante[1] MD, MHS; Aaron Blau[1] MD; Adrian N Berg[1] BS; Aaron S Weinberg[2] MD, MPhil; Ik E Jun[2] MD; Victor F Tapson[2] MD; Tanya S Kanigan[3] PhD; Artur Adib[3] PhD

[1]Larner College of Medicine at the University of Vermont Burlington US
[2]Cedars-Sinai Medical Center Los Angeles US
[3]Biocogniv, Inc South Burlington US

Corresponding Author:
Timothy B Plante MD, MHS
Larner College of Medicine at the University of Vermont
89 Beaumont Ave
Burlington
US

Abstract

Background: Conventional diagnosis of COVID-19 with polymerase chain reaction (PCR) testing is associated with prolonged time to diagnosis and costs of running the test. The SARS-CoV-2 virus might lead to characteristic patterns in levels of widely-available, routine blood tests results that could be identified with machine learning methodologies. Machine learning modalities integrating findings from these common laboratory test results might accelerate ruling out emergency department patients for COVID-19.

Objective: We sought to develop and externally validate a machine learning model to rule out COVID-19 using only routine blood tests among adults in emergency departments.

Methods: Using clinical data from emergency departments (EDs) from 66 US hospitals before the pandemic (before the end of December 2019) or during the pandemic (March-July 2020), we included patients aged ?20 years in the study timeframe or missing laboratory results. Model development used 2,183 PCR-confirmed positive cases from 43 hospitals during the pandemic as positive controls; negative controls were 10,000 pre-pandemic patients from the same hospitals. External validation used 23 hospitals with 1,020 pandemic PCR-positive cases and 171,734 pre-pandemic negative controls. The main outcome was COVID-19 status predicted using same-day routine laboratory results. Model performance was assessed with area under the receiving operating characteristic curve (AUROC) as well as sensitivity, specificity, negative predictive value (NPV).

Results: Of 184,937 patients included (median [IQR] age deciles 50.0 [30.0-60.0] years, 40.5% male), AUROC for development and external validation was 0.91 (95% CI, 0.90-0.92). Using a risk score cutoff of 1.0 (out of 100) in the validation dataset, the model achieved sensitivity of 95.9% and specificity of 41.7%; with a cutoff of 2.0, sensitivity of 92.5% and specificity of 60%. At the cutoff of 2, the NPVs at prevalences of 1%, 10%, and 20% were 99.9%, 98.6%, and 97%.

Conclusions: A machine learning model developed with multicenter clinical data integrating commonly collected ED laboratory data demonstrated high rule-out accuracy for COVID-19 status, and might inform selective use of PCR-based testing. Clinical Trial: N/A

(JMIR Preprints 01/09/2020:24048)
DOI: https://doi.org/10.2196/preprints.24048

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study

Running title: Development & external validation of a COVID-19 rule-out model

Timothy B Plante, MD MHS[a,b], Aaron Blau, MD[b], Adrian N Berg, BS[a,d], Aaron S Weinberg, MD MPhil[c], Ik C Jun MD[c], Victor F Tapson, MD[c], Tanya S Kanigan, PhD[d], Artur B Adib, PhD[d]

[a] Larner College of Medicine at the University of Vermont, Burlington, VT, USA
[b] University of Vermont Medical Center, Burlington, VT, USA
[c] Cedars-Sinai Medical Center, Los Angeles, CA, USA
[d] Biocogniv Inc, South Burlington, VT, USA

Corresponding Author:
Timothy B. Plante, MD MHS
Larner College of Medicine at the University of Vermont
360 S Park Drive
Suite 206B
Colchester, VT 05446
Timothy.Plante@uvm.edu
Telephone: 802-656-3688
Fax: 802-656-8965

Word count:
Abstract: 346 words (≤450 words)
Manuscript: 4732 words
4 Figures/tables: 2 figures, 2 tables

Key words: COVID-19, SARS-CoV-2, machine learning, artificial intelligence, electronic medical records, laboratory results

Abstract

Background

Conventional diagnosis of COVID-19 with reverse transcription polymerase chain reaction (RT-PCR) testing (hereafter, PCR) is associated with prolonged time to diagnosis and costs of running the test. The SARS-CoV-2 virus might lead to characteristic patterns in levels of widely-available, routine blood tests results that could be identified with machine learning methodologies. Machine learning modalities integrating findings from these common laboratory test results might accelerate ruling out emergency department patients for COVID-19.

Objective

We sought to develop (ie, train and internally validate with cross-validation techniques) and externally validate a machine learning model to rule out COVID-19 using only routine blood tests among adults in emergency departments.

Methods

Using clinical data from emergency departments (EDs) from 66 US hospitals before the pandemic (before the end of December 2019) or during the pandemic (March-July 2020), we included patients aged ≥20 years in the study timeframe. We excluded those with missing laboratory results. Model training used 2,183 PCR-confirmed positive cases from 43 hospitals during the pandemic as positive controls; negative controls were 10,000 pre-pandemic patients from the same hospitals. External validation used 23 hospitals with 1,020 pandemic PCR-positive cases and 171,734 pre-pandemic negative controls. The main outcome was COVID-19 status predicted using same-day routine laboratory results. Model performance was assessed with area under the receiver operating characteristic (AUROC) curve as well as sensitivity, specificity, negative predictive value (NPV).

Results

Of 192,779 patients included in the training, external validation, and sensitivity datasets (median [IQR] age deciles 50.0 [30.0-60.0] years, 40.5% [78,249/192,779] male), AUROC for training and external validation was 0.91 (95% CI, 0.90-0.92). Using a risk score cutoff of 1.0 (out of 100) in the external validation dataset, the model achieved sensitivity of 95.9% and specificity of 41.7%; with a cutoff of 2.0, sensitivity of 92.6% and specificity of 59.9%. At the cutoff of 2, the NPVs at prevalences of 1%, 10%, and 20% were 99.9%, 98.6%, and 97%.

Conclusions

A machine learning model developed with multicenter clinical data integrating commonly collected ED laboratory data demonstrated high rule-out accuracy for COVID-19 status, and might inform selective use of PCR-based testing.

Introduction

SARS-CoV-2 is the cause of COVID-19, which continues to spread in an uncontrolled manner across the United States.[1] COVID-19 management includes patient isolation and supportive care.[2] This strategy requires expeditious COVID-19 diagnosis, but components required for the reverse transcription polymerase chain reaction (RT-PCR, hereafter, PCR) assay have been reported to be in short supply in some locations during the pandemic, leading to delays in results.[3] In the absence of a widely available PCR test with rapid turnaround there is an urgent need to identify alternative means for stratifying risk of patients seeking care during the COVID-19 pandemic.

Risk assessment models might identify those at low risk of active COVID-19 using available data from the clinical encounter.[4,5] In contrast to traditional model building techniques, machine learning technologies consider complex linear and non-linear associations between independent variables and identify characteristic patterns of commonly-collected data among patients with COVID-19.[6] A test with high sensitivity and diagnostic yield (ie, fraction of patients ruled out) could be used in a manner analogous to other rule-out tests, such as D-dimer for pulmonary embolism.[7]

Using emergency department (ED) patient encounters from a well-established multicenter clinical database, we sought to 1) describe the development of a machine learning model for ruling out COVID-19 using only routinely-collected laboratory tests, and 2) assess the area under the receiver operating characteristic (AUROC) curve of a machine learning model's concordance with both PCR COVID-19 molecular test results (for positives) and pre-pandemic patients (for negatives). We hypothesized that such a machine learning model would enable the ruling-out of the disease with great sensitivity (>90%) and diagnostic yield (>50%).

Methods

Study Design and Setting

This analysis and its reporting is compliant with the Standards for Reporting Diagnostic Accuracy Studies (STARD) statement.[8] This cross-sectional study was performed using 3 datasets of deidentified, patient-level electronic medical records of adult patients in an ED. The Premier Healthcare Database (PHD) is a large database of 1,041 US hospitals from all 9 US geographic regions defined by the US Census.[9] At time of writing, 155 hospitals contribute SARS-CoV-2 RNA testing results to the PHD. We separately obtained data from Cedars-Sinai Medical Center (CSMC), an 886-bed academic medical center in Los Angeles, CA, and the Beth Israel Deaconess Medical Center (BIDMC), a 673-bed academic medical center in Boston, MA. Descriptions of these datasets, and an inclusion flow diagram, are described in-depth in the **Supplemental Appendix A**.

Pre-pandemic and Pandemic Timeframes

Two timeframes were used, defined by the date of ED visit during the pre-pandemic (before January 2020) and pandemic (March 2020 through July 2020) dates. January 2020 and February 2020 were not included as there was not widespread monitoring or availability of diagnostic tests for COVID-19 in the United States during this timeframe, even though SARS-CoV-2 community transmission was present in the United States during this time. [10] Clinical encounter data from the PHD were available for the pre-pandemic (January 2019-December 2019) and pandemic (March 2020 through July 2020) timeframes . CSMC data were available for COVID-19 positive patients during the pandemic timeframe only (March-April 2020). BIDMC data were available in an extended pre-pandemic timeframe (2008-2019) only for patients who were admitted through the ED.

Selection of Participants

Eligible patient encounters (hereafter, patients) were adults aged ≥20 years in an ED at an included center during one of the pre-pandemic or pandemic timeframes. Patients were excluded if they were missing a laboratory result included in the model on the day of presentation to the ED or if any of their laboratory results were reported with inappropriate units or incorrect specimen type. Patients were defined as PCR COVID-19 positive (hereafter, PCR-positive) if they had a positive SARS-CoV-2 RNA test on the day of presentation to the ED. We chose PCR rather than antigen positivity to define the cases as PCR is commonly used as the reference standard in COVID-19 diagnosis.[11,12]

Training Population and Definition of COVID-19 Cases and Controls

Training occurred in the PHD database only. The PHD training and external validation sets were split by hospital, and only hospitals that reported COVID positives as well as the required blood tests for the model were included in the analysis (64 total). Of these, 43 hospitals were randomly assigned to the training set, and 21 to the external validation set (hereafter, PHD holdout). Cases came from the pandemic timeframe, and any patients in this timeframe without a positive PCR test were excluded. Contemporary COVID-19 PCR assays have elevated false negative rates, which could lead to mislabeled data and hence to degraded model performance.[13] Because of this, pre-pandemic controls randomly selected from the 43 PHD hospitals in the training set were used in place of PCR-negative patients during the pandemic.

External Validation Populations

The external validation dataset used 3 data sources: 952 PCR-positives and 154,341 pre-pandemic visits from the 21 hospitals in the PHD holdout set; 68 PCR-positive patients from CSMC; and 17,393 pre-pandemic (2008-2019) patient encounters from BIDMC. Patients in the pandemic timeframe without a positive PCR test were excluded. All pre-pandemic patients were treated as negatives when evaluating the performance of the model in predicting COVID-19 status. The pre-pandemic patients from the PHD holdout were chosen so as to match the top 20 most frequent primary diagnoses given to non-COVID patients during the pandemic, as coded by Clinical Classifications Software Refined (CCSR) codes (listed in the **Supplemental Appendix B**).

Sensitivity Analysis Population

To evaluate how the model generalizes to pandemic-timeframe patients only, we performed a sensitivity analysis using patients presenting to the ED in the 21 centers from the PHD holdout with any SARS-CoV-2 PCR result available on day of presentation. This differed from the other analyses as negatives were from the same timeframe as the positives. This resulted in a total of 952 PCR-positive patients and 6,890 PCR-negative patients in the pandemic period (March-July 2020).

Subgroup Analyses

The AUROC was tabulated by decile of age; and by sex, race, admission or discharge status, and intensive care unit (ICU) admission status in the external validation dataset. The distribution of risk scores was also visualized for all studied cohorts through boxplots. For PCR-positives, this included positives from CSMC, and PHD visits that had a single positive PCR result as well as visits that had a negative result before a positive result (both on the day of presentation). For PCR negatives during the pandemic, this included patients with both single- and double-PCR results on the day of presentation. For pre-pandemic encounters, the scores for all eligible BIDMC patients were considered, as well as those from the PHD holdout that matched the top 20 CCSR (non-COVID) codes observed during the pandemic.

Model development (ie, training and internal validation with cross-tabulation techniques)

The model was intended to estimate COVID-19 status on the day of presentation to an ED using common laboratory tests collected that day. Model training began with 29 routinely measured features (ie, potential or included model covariates) comprising the comprehensive metabolic panel and the complete blood count with differential. Recursive feature elimination with cross-validation (RFECV) was performed to arrive at the final 15 features.[14] We used the gradient boosting model as implemented in XGBoost [15] for all results. No hyperparameter optimization was performed and default parameters were used. Performance on the training set was evaluated through stratified 5-fold cross-validation. Performance in the external validation and sensitivity analysis datasets were obtained after training the model on the entire training set.

Statistical Analysis

Baseline demographics, ED disposition, and included laboratory features from the training, external validation, and sensitivity analysis datasets were tabulated by COVID-19 status. Visualization of the distribution of features used box plots, ordered by feature importance (cf. **Supplemental Appendix C** lists values). Model discrimination was visualized with receiver operating characteristic (ROC) curves and estimation of the AUROC. AUROC 95% CIs were estimated with bootstrapping. Hosmer-Lemeshow criteria were used to describe performance of discrimination.[16] These criteria considered an AUROC value of 0.5 as no discrimination, 0.5 to <0.7 as poor discrimination, 0.7 to <0.8 as acceptable discrimination, 0.8 to <0.9 as excellent discrimination, and ≥0.9 as outstanding discrimination. Sensitivity, specificity, and negative predictive value (NPV) were defined using conventional definitions. Diagnostic yield was defined as the percentage of patients with a risk score below a given cutoff. All analyses were prespecified. The sample size of this analysis was driven by data availability in this multicenter database.

Analyses were performed in Python v3.7.5 (Python Software Foundation, Beaverton, OR), using the XGBoost package v0.82 [17] and the Scikit-Learn library v0.21.3.[18] The use de-identified database as described here met the non-human subjects research by the University of Vermont's Institutional Review Board criteria.

Results

Demographics and Proportion of PCR-positive in Training Dataset, External Validation Dataset, and Sensitivity Analysis Dataset

The training dataset included 12,183 ED visits at 43 centers from the PHD, of which 2,183 results were PCR-positive. The validation dataset included 172,754 ED visits from 23 centers (21 from the PHD, as well as the independently collected data from CSMC and BIDMC), of which 1,020 results were PCR-positive. The sensitivity analysis dataset included 7,842 records from 21 centers in the PHD holdout group. Patient demographics and visit characteristics are summarized in **Table 1**.

A total of 192,779 eligible patients were included in the study; median (IQR) across age deciles was 50.0 (30.0-60.0) years, 40.5% (78,249/192,779) male. In the training, external validation, and sensitivity analysis datasets, median (IQR) across age deciles was 50.0 (30.0-70.0), 50.0 (30.0-60.0), and 50.0 (40.0-70.0), respectively. Male percentage was 42.9%, 40.1%, and 47.4%, respectively.

Table 1. Demographics of patients and encounter details, by COVID-19 status[1].

	Training (N=12,183)		External Validation (N=172,754)		Sensitivity Analysis (N=7,842)	
COVID-19 Status	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**
n	10,000	2,183	171,734	1,020	6,890	952

Age, n (%)						
20-30 Yrs	1,392 (14)	198 (9)	27,952 (16)	71 (7)	709 (10)	70 (7)
30-40 Yrs	1,481 (15)	304 (14)	29,187 (17)	127 (12)	882 (13)	119 (12)
40-50 Yrs	1,398 (14)	413 (19)	27,764 (16)	214 (21)	896 (13)	205 (22)
50-60 Yrs	1,649 (16)	400 (18)	28,896 (17)	217 (21)	1,172 (17)	208 (22)
60-70 Yrs	1,512 (15)	367 (17)	23,771 (14)	180 (18)	1,200 (17)	163 (17)
70-80 Yrs	1,322 (13)	264 (12)	18,460 (11)	121 (12)	1,063 (15)	108 (11)
80+ Yrs	1,246 (12)	237 (11)	15,704 (9)	90 (9)	968 (14)	79 (8)
Gender, n (%)						
Female	5,876 (59)	1,079 (49)	102,942 (60)	502 (49)	3,650 (53)	477 (50)
Male	4,122 (41)	1,104 (51)	68,790 (40)	518 (51)	3,240 (47)	475 (50)
Unknown	2 (0)	0 (0)	2 (0)	0 (0)	0 (0)	0 (0)
Race, n (%)						
Black	1791 (18)	397 (18)	28,874 (17)	212 (21)	1,230 (18)	201 (21)
Other	904 (9)	976 (45)	23,222 (14)	453 (44)	772 (11)	448 (47)
Unknown	450 (4)	102 (5)	12,284 (7)	48 (5)	368 (5)	36 (4)
White	6,855 (69)	708 (32)	107,354 (63)	307 (30)	4,520 (66)	267 (28)
Census Division[2], n (%)						
East North Central	2,065 (21)	280 (13)	16,184 (9)	108 (11)	1,103 (16)	108 (11)
East South Central	0 (0)	0 (0)	3,549 (2)	50 (5)	138 (2)	50 (5)
Middle Atlantic	782 (8)	294 (13)	18,776 (11)	92 (9)	1,356 (20)	92 (10)
New England	493 (5)	1 (0)	31,624 (18)	1 (0)	1 (0)	1 (0)
Pacific	106 (1)	32 (1)	3,617 (2)	69 (7)	34 (0)	1 (0)
South Atlantic	3,116 (31)	1,192 (55)	70,463 (41)	613 (60)	2,790 (40)	613 (64)
West North Central	633 (6)	39 (2)	0 (0)	0 (0)	0 (0)	0 (0)
West South Central	2,805 (28)	345 (16)	27,521 (16)	87 (9)	1,468 (21)	87 (9)
Rural-Urban[2], n (%)						
Rural	583 (6)	21 (1)	3,617 (2)	1 (0)	34 (0)	1 (0)
Urban	9,417 (94)	2,162 (99)	168,117 (98)	1019 (100)	6,856 (100)	951 (100)
Disposition, n (%)						
Discharge from ED	7,487 (75)	1175 (54)	132,195 (77)	522 (51)	4072 (59)	522 (55)
Non-ICU Admission	2,068 (21)	805 (37)	29,793 (17)	379 (37)	2375 (34)	335 (35)
ICU Admission	445 (4)	203 (9)	9,746 (6)	119 (12)	443 (6)	95 (10)

Abbreviations: ED, emergency department; ICU, intensive care unit; PCR, reverse-transcription polymerase chain reaction; PHD, Premier Healthcare Database.

[1]**For the training dataset**: **COVID-19 positivity** was defined as a positive COVID-19 PCR on the day of presentation to the ED among patients in the pandemic timeframe (March 2020 through July 2020) in the PHD database among a random selection of 43 of the 64 PHD hospitals reporting PCR positives. **COVID-19 negativity** was defined as a selection of 10,000 patients in the pre-pandemic timeframe (January through December 2019) in the PHD database from the same 43 hospitals as the COVID-19 positive patients.

For the external validation dataset: **COVID-19 positivity** was defined the same as for the training dataset for the PHD dataset but also included 952 PCR-positives from the 21 hospitals in the PHD holdout set. Additionally, it included 68 PCR-positive patients from Cedar Sinai Medical Center from March-April of 2020. **COVID-19 negativity** in the external validation set was defined using 154,341 pre-pandemic visits from the 21 hospitals in the PHD holdout set (January through December 2019) and whose primary diagnoses were among the 20 most frequent primary diagnosis given to the COVID-19-negative patients during the pandemic, using Clinical Classification Software Refined codes. It also included 17,393 pre-pandemic (2008-2019) patient encounters from Beth Israel Deaconess Medical Center.

For the sensitivity dataset: **COVID-19 positivity** included the same 952 PCR-positives from the 21 hospitals in the external validation dataset. **COVID-19 negativity** was defined as visits with at least 1 PCR-negative but no PCR-positive results on the day of presentation, and included all 6,890 patients with such results from the same 21 hospitals as the positives.

[2]Census division was defined using US Census classification.[19] Rural areas are considered territory outside of the US Census Bureau's definition of urban. [20] These geographic descriptions pertain to the hospital, not the patient's permanent residence.

Selected Features Included in the Model and Individual Feature Performance

The RFECV method led to the final set of 15 features listed in **Supplemental Table S1**. The distributions of these features in the training dataset, stratified by COVID-19-positive and COVID-19-negative status, ordered by importance to the model are shown in **Supplemental Figure S1**. The largest calculated importance of these features was eosinophils, calcium, and AST. Summary statistics of these features in the training, external validation, and sensitivity analysis datasets stratified by COVID-19 status appears in **Supplemental Table S1.**

Performance of Individual Features and Model Performance in the Training Dataset

The AUROC for each individual feature in the training dataset is shown in **Supplemental Figure S2**. The highest AUROCs were observed for eosinophils, calcium, and aspartate aminotransferase (0.70-0.80). The final model's AUROC in the training dataset was 0.91 (95% CI, 0.90-0.92) (**Figure 1**).

Figure 1. Discrimination as assessed by ROC curves for training, external validation, and sensitivity analysis datasets[1].

Abbreviations: AUROC, area under the receiver operating characteristic curve; PCR, polymerase chain reaction, PHD, Premier Healthcare Database; ROC, receiver operating characteristic.

[1]Receiver operating characteristic (ROC) curves for the 3 different datasets: Training (blue), External Validation (orange), and Sensitivity analysis (green). The training curve was obtained through 5-fold cross-validation, where positive controls are PCR-confirmed cases during the pandemic (N=2,183) and negative controls are pre-pandemic patients (N=10,000) from 43 hospitals in the PHD. The training AUROC was 0.91 (95% CI 0.90-0.92). The external validation curve was performed in the external validation dataset after training the model on the training dataset. External validation positives are PCR-confirmed cases from Cedars-Sinai Medical Center (N=68) and from the PHD holdout set (N=952) comprising 21 hospitals. External validation negatives are pre-pandemic (2019) patients from the same 21 PHD hospitals and matching the top 20 primary non-COVID diagnoses in 2020 (N=154,341), as well as all eligible pre-pandemic (2008-2019) Beth Israel Deaconess Medical Center patients (N=17,393). The AUROC in the external validation dataset was 0.91 (95% CI 0.90-0.92). The sensitivity analysis curve demonstrates the effect of using pre-pandemic patients as negative controls compared to using PCR-negatives from 2020. In this dataset, both positives (N=952) and negatives (N=6,890) are PCR-confirmed patients from the PHD holdout set (21 hospitals), and no pre-pandemic data is included. The AUROC in the sensitivity analysis set was 0.89 (95% CI 0.88-0.90).

Model Performance in the External Validation Dataset

The model's AUROC in the external validation dataset was 0.91 (95% CI, 0.90-0.92), as shown above in **Figure 1**. This corresponds to an outstanding discrimination per the Hosmer-Lemeshow criteria.[16] Sensitivity and specificity, respectively, were 95.9 and 41.7 at a score cutoff of 1, 92.6 and 60.0 at a score of 2, 85.5 and 78.5 at a cutoff of 5, and 79.4 and 87.6 at a cutoff of 10 (**Table 2**).

With a COVID population prevalence of 1%, each of these cutoffs had an NPV >99%; at 10% prevalence, each was >97%, and at a prevalence of 20%, each was >94%. The diagnostic yield ranged from 34% (20% prevalence, score cutoff of 1) to 87% (1% prevalence, score cutoff of 10).

Table 2. Clinical performance metrics for the model in the external validation dataset for various score cutoffs and COVID-19 pretest prevalences[1].

Score Cutoff	Sensitivity	Specificity	Likelihood ratio[2]	Prevalence of 1%		Prevalence of 10%		Prevalence of 20%	
				NPV (%)	Yield[3] (%)	NPV (%)	Yield (%)	NPV (%)	Yield (%)
1	95.9	41.7	0.099	99.9	41.3	98.9	38.0	97.6	34.2
2	92.6	60.0	0.124	99.9	59.4	98.6	54.7	97.0	49.5
5	85.5	78.5	0.185	99.8	77.8	98.0	72.1	95.6	65.7
10	79.4	87.6	0.235	99.8	86.9	97.4	80.9	94.4	74.2

Abbreviations: NPV, negative predictive value.
[1]The maximum score was 100, higher score indicates higher model prediction of COVID-19 positivity.
[2]Likelihood ratio uses the equation for negative tests.
[3]Yield=diagnostic yield, which is the percentage of patients that can be ruled out, ie, with score below the cutoff.

Sensitivity Analysis and Subgroup Analyses

Figure 1, above, depicts the ROC curve in the sensitivity analysis dataset, which contains only year 2020 patients with PCR-confirmed positive and negative results (ie, no historical negatives). The AUROC was 0.89 (95% CI, 0.88-0.90). In **Figure 2** the AUROC is presented for various demographic cohorts as well as patient disposition (ED discharge, non-ICU, and ICU) in the external validation dataset. AUROCs ranged from 0.86 to 0.93.

Figure 2. Discrimination as assessed by AUROC in age, sex, race, and ED disposition subgroups in the external validation dataset[1].

Abbreviations: AUROC, area under the receiver operating characteristic curve; ED, emergency department; ICU, intensive care unit. Non-ICU patients were admitted to the hospital but not to an ICU.
[1]Distribution of AUROCs per demographic, as well as per patient disposition type (ED discharge, non-ICU, and ICU) in the external validation dataset. Top numbers are AUROCs, bottom numbers in parentheses are the number of patients.

Distribution of Risk Scores in Selected Subgroups

Lastly, an extensive distribution of risk scores for various cohorts in the external validation dataset is shown in **Supplemental Figure S3**, including pre-pandemic patients whose primary diagnoses were among the top 20 primary diagnoses among non-COVID patients in 2020. From visual inspection, it can be seen that high scores track PCR-positive patients consistently across all cohorts.

Discussion

Principal Results

A development and external validation study of a machine learning model for COVID-19 status using laboratory tests routinely collected in adult ED patients found high discrimination across age, race, sex, and disease severity subgroups. This model had high diagnostic yield at low score cutoffs in screening population prevalence of 10% or lower. Such a model could rapidly identify those at low risk for COVID-19 in a ‘rule-out’ method, and might reduce the need for PCR testing in such patients.

Comparison with Prior Work

Prior literature has described the application of machine learning techniques to commonly-collected laboratory data for estimation of missing laboratory analytes. For example, an analysis by Waljee and colleagues leveraged machine learning techniques for imputation of missing laboratory data in cohorts of patients with cirrhosis and inflammatory bowel disease at a single institution.[21] In comparison to other common imputation techniques described in this manuscript, the machine learning technique introduced the least imputation error for these laboratory data. Luo and colleagues used similar methods to estimate ferritin from a single medical center, and found the machine learning technique to outperform traditional imputation methods.[22] These serve as strong evidence of the potential use of machine learning for use in estimation of laboratory data. However, outside of imputation of missing values from research databases, the clinical utility for such techniques prior to the current COVID-19 pandemic was unclear.

In the current pandemic, there is an urgent need to rapidly identify patients with the infection to inform supportive clinical care. Prior work has attempted to integrate combinations of clinical data points in diagnostic models, though only a few are currently published in peer-reviewed literature.[23] The selection of the specific points to integrate into machine learning models for COVID-19 diagnosis has implications on its integration into existing clinical delivery. In contrast with the results here, which only included components of the routinely-collected complete blood count with differential and complete metabolic panel laboratory tests, others have integrated non-laboratory features. Sun and colleagues reported 3 models including demographic, radiological, and symptomatology and obtained AUCs ranging from 0.65 to 0.88 for these models.[24] Symptomatology was not obtained with structured, validated questionnaires and the ability to capture these symptoms in a reproducible manner might be difficult outside of a research setting. Further, modern medical records cannot integrate such symptoms into automated risk scores as they are not documented in a structured way.

Structured data obtained routinely in the clinical exam are the simplest to integrate, and might have the least variability between institutions. These include vital signs, demographics, laboratory findings, and radiological images. There are few studies describing the use of such data for diagnosis of COVID-19. One study found a machine learning method to have an accuracy of 87% for distinguishing between COVID-19 from pneumonia or no relevant findings using chest radiographs.[25] A different model developed from chest computed tomography images reported an AUROC of 0.994 when distinguishing between COVID-19 and atypical or viral pneumonias.[26] However, national organizations recommend against the use of radiological imaging for diagnosis of COVID-19, in part because of the added risk of spreading infection through additional visitation to radiology suites. [27] These models are unlikely to be readily deployed because of the challenges of performing elective radiological tests during the present pandemic.

An additional consideration in the development of machine learning models is the inclusion of an adequate sample size for model training.[28-30] Other studies have investigated the role of laboratory data with or without other non-radiological structured clinical data or demographics for the diagnosis of COVID-19 using machine learning techniques. For example, Wu and colleagues reported a C-index of 0.99 but included only 108 patients (12 COVID-19 positive) in their training.[31] Similarly, individual efforts led by Batista, Brinati, and Soares describe machine learning models trained on 234, 279, and 599 patients, respectively.[32-34] These studies are also limited in the small number of centers from which patients were enrolled, and lack of diversity in their patient population.

Advancement of Scientific Knowledge

The present analysis advances science in several key ways. First, we describe a machine learning model developed in a diverse patient population with routine laboratory data from multiple clinical

centers across the United States.[35] Second, the model incorporates common laboratory tests that are widely available with rapid turnaround time. As the machine learning model can be performed essentially instantaneously, the primary time limitation is related to phlebotomy and specimen processing. There is a well-known bottleneck in completing conventional COVID-19 PCR assays; a commercial laboratory recently reported a 7-day reporting lag.[36] Third, the present model could identify those at lowest risk for COVID-19 to inform a 'rule-out' method for screening. Those with intermediate or greater risk for COVID-19 could be further assessed with COVID-19 PCR testing, if indicated. Depending on the selected score cutoff and population prevalence, such an approach could rule out 34% to 87% of ED patients requiring conventional COVID-19 PCR testing (see yield, **Table 2**). The specific score cutoff for rule out of COVID-19 with this model can be customized based upon what an institution considers to be an 'acceptable' target NPV. However, the diagnostic yield will change based upon the screening population prevalence of COVID-19, and the diagnostic yield will be inversely related to screening population prevalence of COVID-19. For example, assume that an institution determines that an acceptable NPV for this model is 97.5%. If this institution's screening population has a 20% prevalence of COVID-19, the threshold score cutoff would be set at 1, and the diagnostic yield (i.e., the percentage of patients ruled out for COVID-19 at a score cutoff of 1) would be 34.2% (see **Table 2**). However, at a prevalence of 10%, the score cutoff threshold would be 10, and the diagnostic yield would be 80.9%. The efficiency of diagnostic yield with this model is higher at lower prevalence. Finally, the sensitivity of the present model at a score cutoff of 1, 2, and 5 (95.9, 92.6, 85.5, respectively) was similar to COVID-19 antigen assays (66.1 to 86.3) and sputum and saliva PCR assays (62.3 to 97.2).[11] The comparatively similar sensitivities between the model and these exiting assays supports the clinical utility of machine learning models as future diagnostic tools. .

Weaknesses and Strengths

This study has weaknesses. While the choice of pre-pandemic controls partially circumvents the issue of false negatives in PCR testing by ensuring the negatives that the model is trained on are true negatives, it does not ensure that the positives encompass the full spectrum of true positives, since those are sometimes missed by PCR due to changes in viral load as a function of disease progression.[37,38] Additionally, the use of controls from a different time period could introduce a bias of its own, such as different demographics or non-COVID morbidities. However, the sensitivity analysis used COVID-19 positive and negatives from the pandemic timeframe, and performance of the model was reassuringly similar to the performance in the external validation. The performance in demographic, clinical diagnosis, and ED disposition subgroups were also similar to the external validation. Laboratory data were performed locally at each hospital, not centrally performed. The model requires all components of the laboratory data to be included. This study only included patients who visited an ED. While it is likely that some of the patients in this study were asymptomatic or presymptomatic and were found to have COVID-19 as part of routine admission, we were unable to determine the indications for screening and therefore are unable to determine the performance of this model in asymptomatic and presymptomatic adults. The present analysis only accounted for results from COVID-19 PCR and not for alternative diagnostic methods, such as antigen for acute infection or antibody testing to demonstrate prior infection. Finally, the research database did not include details about the specific PCR assay used in diagnosis, so we are unable to comment on performance of the model in comparison to the performance of the specific assays.

This study has strengths. This study included data from a large number of patients and hospitals, and to our knowledge is the largest application of machine learning to COVID-19. Data were derived from an electronic medical records database that is commonly used in clinical research. The patient population was geographically and racially diverse. The only features included in the model were from blood tests that are already routinely collected in ED encounters. Further, these tests were from

multiple hospitals, suggesting that the model is robust against different specimen collection, handling practices, and instrumentation. Sensitivity analyses were performed to evaluate potential biases due to the choice of pre-pandemic negative controls, and no significant bias was observed across multiple dimensions. Our methods extend on established machine learning-based imputation methods for missing laboratory data [21,22], and suggests there may be clinical utility of these techniques in disease rule out. Finally, the external validation was a true external validation since it used data from hospitals that were not included in the training dataset. This supports the resilience of the model across institutions with differing specimen handling and laboratory processing methods.

Conclusions

A machine learning model for ruling out COVID-19 in ED patients that integrates commonly-collected laboratory data had a discrimination accuracy that can be classified as excellent to outstanding.[16] Using score cutoffs of 5 and 10 points, and assuming a 10% screening population prevalence of COVID-19, 72% and 81% of patients were ruled out with this model, respectively, while maintaining an NPV >97%.

Acknowledgments

We thank Alistair EW Johnson, PhD and John House MS for providing technical expertise related to the BIDMC and PHD datasets respectively. We would also like to thank Ryan McGinnis, PhD, who provided helpful feedback on the manuscript, and Dr. Robert A. Levine, MD, R. George Hauser, MD, and Mark K. Fung, MD for various answers to laboratory medicine questions. We thank Susan Gifford for her helpful review of this manuscript.

Conflicts of Interest

Disclosures: TBP, AB, ASW, and ICJ have no disclosures. VFT received research funding from the National Institutes of Health (paid to the institution) for COVID-related clinical trials. ANB is a paid intern at Biocogniv. TSK and ABA have ownership of Biocogniv.

References

1. Coronavirus in the U.S.: Latest map and case count - The New York Times. The New York Times Web site. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html. Published 2020. Accessed August 4, 2020.
2. Baden LR, Rubin EJ. Covid-19 - the search for effective therapy. N Engl J Med. 2020;382(19):1851-1852. doi:10.1056/NEJMe2005477. PMID:32187463.
3. Karow J. Scientists seek solution to coronavirus testing bottleneck. Crain Communications. https://www.modernhealthcare.com/patients/scientists-seek-solution-coronavirus-testing-bottleneck. Published 2020. Accessed August 4, 2020.
4. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38(23):1805-1814. doi:10.1093/eurheartj/ehw302.
5. Peterson ED. Machine learning, predictive analytics, and clinical practice: can the past inform the present? JAMA. 2019. doi:10.1001/jama.2019.17831. PMID:31755902.
6. Waljee AK, Higgins PDR. Machine learning in medicine: a primer for physicians. Am J Gastroenterol. 2010;105(6):1224-1226. doi:10.1038/ajg.2010.173. PMID:20523307.
7. Kearon C, de Wit K, Parpia S, et al. Diagnosis of pulmonary embolism with d-dimer adjusted to clinical probability. N Engl J Med. 2019;381(22):2125-2134. doi:10.1056/NEJMoa1909159. PMID:31774957.
8. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem. 2015;61(12):1446-1452. doi:10.1373/clinchem.2015.246280. PMID:26510957.
9. Premier Healthcare Database (COVID-19) white paper: Data that informs and performs. Premier Inc.; 2020/04/10/ 2020. https://products.premierinc.com/downloads/PremierHealthcareDatabaseWhitepaper.pdf
10. CDC. Evidence for limited early spread of COVID-19 within the United States, January-February 2020. MMWR Morb Mortal Wkly Rep. 2020;69(22):680-684. doi:10.15585/mmwr.mm6922e1.
11. Boger B, Fachi MM, Vilhena RO, Cobre AF, Tonin FS, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. Am J Infect Control. 2020. doi:10.1016/j.ajic.2020.07.011. 32659413.
12. Axell-House DB, Lavingia R, Rafferty M, Clark E, Amirian ES, Chiao EY. The estimation of diagnostic accuracy of tests for COVID-19: A scoping review. J Infect. 2020. doi:10.1016/j.jinf.2020.08.043. 32882315.
13. Redman T. If your data is bad, your machine learning tools are useless. Harvard Business School Publishing. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Published 2018. Accessed August 5, 2020.
14. scikit-learn. sklearn.feature_selection.RFECV — scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html. Published 2020. Accessed August 1, 2020.
15. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Paper presented at: 22nd ACM SIGKDD International Conference; 2016/08/13/, 2016. doi:10.1145/2939672.2939785.
16. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013. doi:10.1002/9781118548387

17. XGBoost. XGBoost GPU Support — xgboost 0.82 documentation. https://xgboost.readthedocs.io/en/release_0.82/gpu/index.html. Published 2016. Accessed August 4, 2020.
18. Albon C. Machine learning with Python cookbook: practical solutions from preprocessing to deep learning. O'Reilly Media; 2018 (1st edition). https://www.amazon.com/Machine-Learning-Python-Cookbook-Preprocessing/dp/1491989386. Accessed December 11, 2019.
19. USCB. 2010 census regions and divisions of the United States. U.S. Census Bureau. https://www.census.gov/geographies/reference-maps/2010/geo/2010-census-regions-and-divisions-of-the-united-states.html. Updated August 20, 2018. Accessed August 4, 2020.
20. USCB. 2010 census urban and rural classification and urban area criteria. United States Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html. Updated December 2, 2019. Accessed August 4, 2020.
21. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8). doi:10.1136/bmjopen-2013-002847. PMID:23906948.
22. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol. 2016;145(6):778-788. doi:10.1093/ajcp/aqw064. PMID:27329638.
23. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi:10.1136/bmj.m1328. PMID:3226522.
24. Sun Y, Koh V, Marimuthu K, et al. Epidemiological and clinical predictors of covid-19. Clin Infect Dis. 2020;71(15):786-792. doi:10.1093/cid/ciaa322. PMID:32211755.
25. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020;121:103792. doi:10.1016/j.compbiomed.2020.103792. PMID:32568675.
26. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med. 2020;121:103795. doi:10.1016/j.compbiomed.2020.103795. PMID:32568676.
27. ACR. ACR recommendations for the use of chest radiography and computed tomography (CT) foR SUSPECTED COVID-19 infection | american college of radiology. American College of Radiology. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed August 5, 2020.
28. Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods. 2015;250:126-136. doi:10.1016/j.jneumeth.2015.01.010.
29. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell. 1991;13(3):252-264. doi:10.1109/34.75512.
30. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365. PMID:31697686.
31. Wu J, Zhang P, Zhang L, et al. Rapid and accurate identification of COVID-19

infection through machine learning based on clinical available blood test results. medRxiv. 2020. doi:10.1101/2020.04.02.20051136.
32. Batista AFdM, Miraglia JL, Donato THR, Chiavegatto Filho ADP. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. 2020. doi:10.1101/2020.04.04.20052092.
33. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. J Med Syst. 2020;44(8):135. doi:10.1007/s10916-020-01597-4.
34. Soares F, Villavicencio A, Anzanello M, Fogliatto F, Idiart M, Stevenson M. A novel high specificity COVID-19 screening method based on simple blood exams and artificial intelligence. medRxiv. https://covid-19.conacyt.mx/jspui/handle/1000/3581. Published 2020. Accessed August 5, 2020.
35. Rehmani R, Amanullah S. Analysis of blood tests in the emergency department of a tertiary care hospital. Postgrad Med J. 1999;75(889):662-666. doi:10.1136/pgmj.75.889.662. PMID:10621876.
36. Picchi A. Quest Diagnostics says COVID-19 test turnaround now 7 days due to surge - CBS News. CBS Interactive Inc. https://www.cbsnews.com/news/quest-diagnostics-coronavirus-testing-seven-day-turnaround-case-surge/. Accessed August 3, 2020.
37. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. Ann Intern Med. 2020. doi:10.7326/M20-1495. PMID:32422057.
38. Woloshin S, Patel N, Kesselheim AS. False negative tests for SARS-CoV-2 infection - challenges and implications. N Engl J Med. 2020;383(6):e38. doi:10.1056/NEJMp2015897. PMID:32502334.

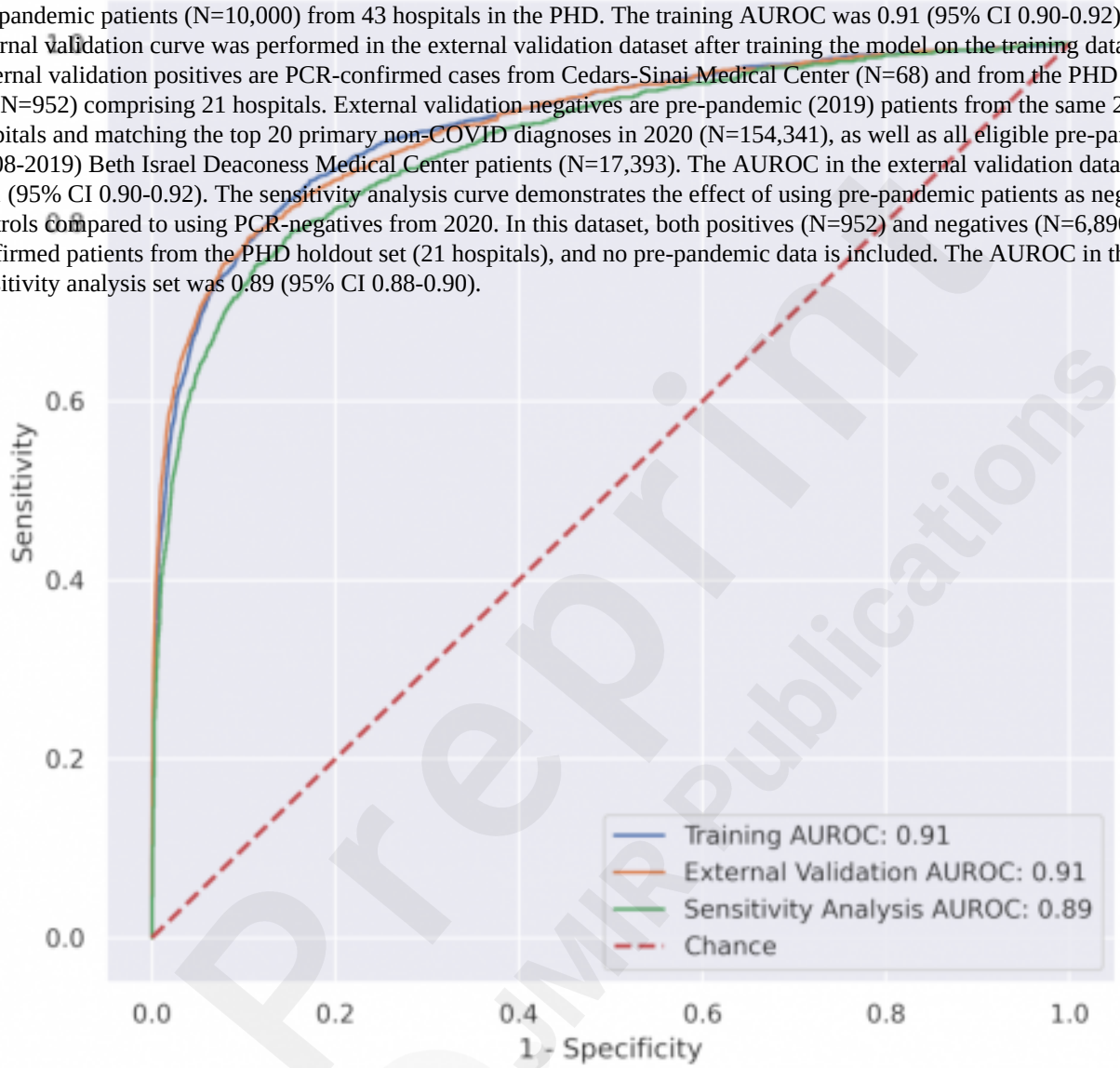
Abbreviations

AUC: area under the curve
AUROC: area under the receiver operating characteristic
BIDMC: Beth Israel Deaconess Medical Center
CCSR: Clinical Classifications Software Refined
CSMC: Cedars-Sinai Medical Center
ED: emergency department
ICU: intensive care unit
NPV: negative predictive value
PCR: reverse transcription polymerase chain reaction
PHD: Premier Healthcare Database
RFECV: recursive feature elimination with cross-validation
ROC: receiver operating characteristic
STARD: Standards for Reporting Diagnostic Accuracy Studies

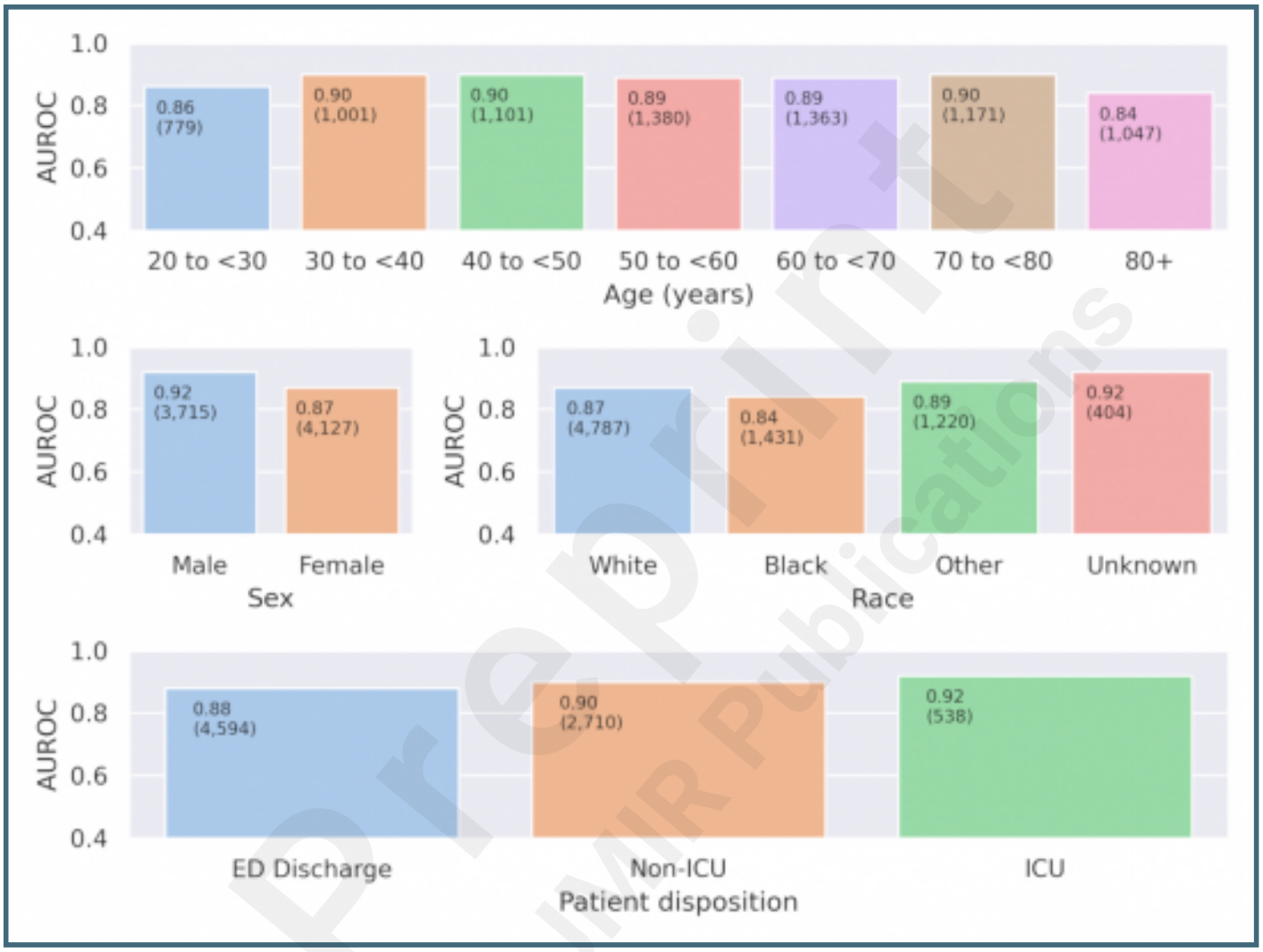
Supplementary Files

Figures

Discrimination as assessed by ROC curves for training, external validation, and sensitivity analysis datasets. Abbreviations: AUROC, area under the receiver operating characteristic curve; PCR, polymerase chain reaction, PHD, Premier Healthcare Database; ROC, receiver operating characteristic. Receiver operating characteristic (ROC) curves for the 3 different datasets: Training (blue), External Validation (orange), and Sensitivity analysis (green). The training curve was obtained through 5-fold cross-validation, where positive controls are PCR-confirmed cases during the pandemic (N=2,183) and negative controls are pre-pandemic patients (N=10,000) from 43 hospitals in the PHD. The training AUROC was 0.91 (95% CI 0.90-0.92). The external validation curve was performed in the external validation dataset after training the model on the training dataset. External validation positives are PCR-confirmed cases from Cedars-Sinai Medical Center (N=68) and from the PHD holdout set (N=952) comprising 21 hospitals. External validation negatives are pre-pandemic (2019) patients from the same 21 PHD hospitals and matching the top 20 primary non-COVID diagnoses in 2020 (N=154,341), as well as all eligible pre-pandemic (2008-2019) Beth Israel Deaconess Medical Center patients (N=17,393). The AUROC in the external validation dataset was 0.91 (95% CI 0.90-0.92). The sensitivity analysis curve demonstrates the effect of using pre-pandemic patients as negative controls compared to using PCR-negatives from 2020. In this dataset, both positives (N=952) and negatives (N=6,890) are PCR-confirmed patients from the PHD holdout set (21 hospitals), and no pre-pandemic data is included. The AUROC in the sensitivity analysis set was 0.89 (95% CI 0.88-0.90).



Discrimination as assessed by AUROC in age, sex, race, and ED disposition subgroups in the external validation dataset. Abbreviations: AUROC, area under the receiver operating characteristic curve; ED, emergency department; ICU, intensive care unit. Non-ICU patients were admitted to the hospital but not to an ICU. Distribution of AUROCs per demographic, as well as per patient disposition type (ED discharge, non-ICU, and ICU) in the external validation dataset. Top numbers are AUROCs, bottom numbers in parentheses are the number of patients.



Multimedia Appendixes

Supplementary figures and tables.
URL: https://asset.jmir.pub/assets/76b25590a33729479d0c0cf96e8c541d.docx

CONSORT (or other) checklists

STARD Checklist.
URL: https://asset.jmir.pub/assets/4fed4bed8f727e68d656f6cb9b9f8592.pdf