Development and validation of the COVID-19 severity index (CSI): a prognostic tool for early respiratory decompensation

Adrian Haimovich MD PhD[1,*]; Neal G. Ravindra PhD[2,3,*]; Stoytcho Stoytchev MS[1]; H. Patrick Young PhD[4,5]; F. Perry Wilson[5,6]; David van Dijk PhD[2,3]; Wade L. Schulz MD PhD[4,7,8]; R. Andrew Taylor MD MHS[1,7,†]

[1]Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT; [2]Cardiovascular Research Center, Section of Cardiovascular Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT; [3]Department of Computer Science, Yale University, New Haven, CT; [4]Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT; [5]Department of Internal Medicine, Yale University School of Medicine, New Haven, CT; [6]Clinical and Translational Research Accelerator, Department of Medicine, Yale University School of Medicine, New Haven CT; [7]Center for Medical Informatics, Yale University School of Medicine, New Haven, CT; [8]Department of Laboratory Medicine, Yale University School of Medicine, New Haven, CT [9]Department of Computer Science, Yale University, New Haven, CT;

Abstract

Objective: The goal of this study was to create a predictive model of early hospital respiratory decompensation among patients with COVID-19.

Design: Observational, retrospective cohort study.

Setting: Nine-hospital health system within the Northeastern United States.

Populations: Adult patients ($\geq$ 18 years) admitted from the emergency department who tested positive for SARS-CoV-2 (COVID-19) up to 24 hours after initial presentation. Patients meeting criteria for critical respiratory illness within 4 hours of arrival were excluded.

Main outcome and performance measures: We used a composite endpoint of respiratory critical illness as defined by oxygen requirement beyond low-flow nasal cannula (e.g., non-rebreather mask, high-flow nasal cannula, bi-level positive pressure ventilation), intubation, or death within the first 24 hours of hospitalization. We developed predictive models using patient demographic and clinical data collected during those first 4 hours. Eight hospitals were used for development and internal validation ($n = 932$) and 1 hospital for model external validation ($n = 240$). Predictive variables were identified using an ensemble approach that included univariate regression, random forest, logistic regression with LASSO, Chi-square testing, gradient boosting information gain, and gradient boosting Shapley additive explanation (SHAP) values prior to manual curation. We generated two predictive models, a quick COVID-19 severity index (qCSI) that uses only exam and vital sign measurements, and a COVID-19 severity index (CSI) machine learning model. Using area under receiver operating characteristic (AU-ROC), precision-recall curves (AU-PRC) and calibration metrics, we compare the qCSI and CSI to three illness scoring systems: Elixhauser mortality score, qSOFA, and CURB-65. We present performance of qCSI and CSI on an external validation cohort.

Results: During the study period from March 1, 2020 to April 27, 2020, 1,792 patients were admitted with COVID-19. Six-hundred and twenty patients were excluded based on age or critical illness within the first 4 hours, yielding 1172 patients in the final cohort. Of these patients, 144 (12.3%) met the composite endpoint within the first 24 hours. The qCSI (AU-ROC: 0.90 [0.85-0.96]) comprised of nasal cannula flow rate, respiratory rate, and minimum documented pulse oximetry outperformed the baseline models (qSOFA: 0.76 [0.69-0.85]; Elixhauser: 0.70 [0.62-0.80]; CURB-65: AU-ROC 0.66 [0.58-0.77]) and was validated on an external cohort (AU-ROC: 0.82). The machine learning-based CSI had superior performance on the training cohort (AU-ROC: 0.91 [0.86-0.97]), but was unlikely to provide practical improvements in clinical settings.

Conclusions: A significant proportion of admitted COVID-19 patients decompensate within 24 hours of hospital presentation and these events are accurately predicted using respiratory exam findings within a simple scoring system.

Introduction

The SARS-CoV-2 disease (COVID-19) is increasingly understood to be a disease with a significant rate of critical illness. International reports of intensive care unit (ICU) utilization frequencies have varied from less than 10% to above 30%.[1–3] There are now reports from larger ICU cohorts, but these do not report a denominator of total COVID-19 population.[4,5] More recently, a large New York City, USA case series was presented, of which 14.2% of patients with known outcomes were admitted to the ICU.[6] Preliminary data from a second New York City, USA cohort had an ICU rate of 32.5%.[7]

While there is a growing body of data about critically ill cohorts and outcomes, less is known about risk factors for critical illness, especially as they relate to respiratory status. Oxygen saturation and inflammatory markers including d-dimer, ferritin, and C-reactive protein (CRP) have been identified as potentially associated with critical illness.[7] Predictive models advance the purposes of risk factor analysis and, ideally, lay the groundwork for the assignment of individualized illness probabilities. A number of diagnostic and prognostic prediction models for COVID-19 have been proposed, but the included cohorts were small and at significant risk for bias.[8]

In this work, we expand on previous efforts describing critical COVID-19 illness in three ways. First, we describe the prevalence of patient respiratory deterioration early ($<$ 24 hours) during hospitalization. While clinical decompensation can occur at any point during a hospitalization, we focus on early escalations in oxygen requirements, which have significant implications for resource utilization and anticipatory guidance for patients and families. Of particular note is the need for urgent patient re-evaluation of patients on general medical wards in consideration of higher levels of care. This process is personnel intensive, often including ward providers, a rapid response team, and intensive care consultants, and can lead to use of multiple care areas at a time when hospital censuses are already stretched.[9,10]

Second, to aid healthcare providers in assessing illness severity in COVID-19 positive patients, we present two predictive models of early respiratory decompensation during hospitalization: the quick COVID-19 severity index (qCSI) and a machine learning-derived COVID-19 Severity Index (CSI). These models were built on data extracted from the first four hours of care. We compare the predictive capabilities of our model to three benchmarks accessible using data in our electronic health record: the Elixhauser comorbidity mortality score,[11] the quick sequential organ failure assessment (qSOFA)[12,13], and the CURB-65 pneumonia severity score.[14] While many clinical risk models exist, these benefit from wide clinical acceptability and relative model parsimony as they require minimal input data for calculation. The Elixhauser comorbidity score was derived to enable prediction of hospital death using administrative data.[11] The qSOFA score was included in SEPSIS-3 guidelines and can be scored at the bedside as it includes respiratory rate, mental status, and systolic blood pressure.[12] The CURB-65 pneumonia severity score has been well-validated for hospital disposition, but its utility in both critical illness and COVID-19 is, as of yet, unclear.[14,15]

Third, we make the qCSI available to the public via a web interface at covidseverityindex.org. This web portal hosts the parsimonious model and allows for user entry of the required clinical values.

Methods

Study Design and setting

This was an observational study to develop a prognostic model of early respiratory decompensation in patients admitted from the emergency department with COVID-19. The healthcare system is comprised of a mix of pediatric ($n = 1$), suburban community ($n = 6$), urban community ($n = 2$), and urban academic ($n = 1$) emergency departments. Data from eight hospitals were used in the creation and internal validation of the predictive model, while data from the last site was withheld for external validation. We adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and STROBE checklists.[16,17]

Data collection and processing

Patient demographics, summarized past medical histories, vital signs, outpatient medications, chest x-ray (CXR) reports, and laboratory results available during the ED encounter were extracted from our local Observational Medical Outcomes Partnership data repository and analyzed within our computational health platform.[18] Data were collected

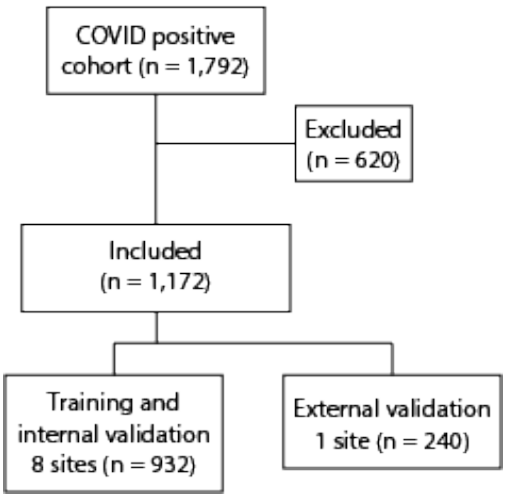


Figure 1: Study Flow Diagram

into a research cohort using custom scripts in PySpark (version 2.4.5) that were reviewed by an independent analyst.

Non-physiologic values likely related to data entry errors for vitals were converted to missing values based on expert-guided rules (available in supplemental code files). Laboratory values at minimum or maximum thresholds and encoded with "<" or ">" were converted to the numerical threshold value and other non-numerical values were dropped. Past medical histories were generated by using diagnoses prior to the date of admission to exclude new diagnoses. Outpatient medications were mapped to their respective First DataBank Enhanced therapeutic classification system.[19] CXR reports were manually reviewed by two physicians and categorized as "no opacity", "unilateral opacity", or "bilateral opacities". One hundred x-ray reports were reviewed by both physicians to determine inter-rater agreement with weighted kappa.

Critical respiratory illness determination

We define critical respiratory illness in the setting of COVID-19 as any COVID-19 patient meeting one of the following criteria: low-flow oxygenation greater than or equal to 10 liters by nasal cannula, high-flow oxygenation, noninvasive ventilation, invasive ventilation, or death. At the start of the COVID-19 pandemic, ICU admissions within our health system were protocolized to include low-flow nasal cannula with intensivist consultation. Since this practice has since evolved, we do not include intensive care unit admission in our composite outcome. A subset of outcomes were manually reviewed by physician members of the institutional computational healthcare team as part of a system wide process to standardize outcomes for COVID-19 related research.

Inclusion and exclusion criteria

Data included visits from March 1, 2020 through April 27, 2020 as our institution's first COVID-19 tests were ordered after March 1, 2020. This study included COVID-19 positive patients as determined by test results ordered between 14 days prior to and up to 24 hours after hospital presentation. We included delayed testing because institutional guidelines initially restricted testing within the hospital to inpatient wards. Testing for COVID-19 was performed at local and/or reference laboratories by nucleic acid detection methods using oropharyngeal (OP), nasopharyngeal (NP), or a combination OP/NP swab. We excluded patients less than 18 years of age and those who met our critical illness criteria at any point within four hours of presentation. The latter of these criteria was intended to exclude patients for whom critical illness was nearly immediately apparent to the medical provider and for whom a prediction would not be helpful. Patients who explicitly opted out of research were excluded from analysis ($n < 5$). Twenty-four hour outcomes for all patients were extracted from the electronic health record.

Baseline models

We generated comparator models using Elixhauser comorbidity mortality scores, qSOFA, and CURB-65. ICD-10 codes from patient past medical histories were mapped to Elixhauser comorbidity groups and mortality scores using H-CUP Software and Tools (*hcuppy* package, version 0.0.7).[20,21]. Where multiple vital signs were available, the worst value was used in score calculation (e.g., the lowest recorded systolic blood pressure for qSOFA). Where no Glasgow Coma Scale (GCS) was recorded, a normal mental status (GCS = 15) was assumed. qSOFA was calculated as the sum of the following findings, each of which were worth one point: GCS $<$ 15, respiratory rate $\geq$ 22, and systolic blood pressure $\leq$ 100. CURB-65 was calculated as the sum of the following findings, each of which were worth one point: GCS $<$ 15, BUN $>$ 19 mg/dL respiratory rate $\geq$ 30, systolic blood pressure $<$ 90 mmHg or diastolic $\leq$ 60 mmHg, and age $\geq$ 65 years.

Severity indices

Samples from eight hospitals were used in model generation and internal validation with the remaining large, urban community hospital serving as an independent test set for external validation of the CSI. All models were were fit on patient demographic and clinical data collected during the first 4 hours of patient presentation. We used an ensemble technique to identify and rank potentially important predictive variables based on their occurrence across multiple selection methods: univariate regression, random forest, logistic regression with LASSO, Chi-square testing, gradient boosting information gain, and gradient boosting Shapley additive explanation (SHAP) interaction values.[22–24] We counted the co-occurences of the the top 20, 30, and 40 variables of each of the methods prior to selecting features for a minimal scoring model (qCSI) and machine learning model (CSI) using gradient boosting. For the qCSI, we used a point system guided by logistic regression. The gradient-boosting CSI model was fit using the XGBoost package and hyperparameters were set using a Bayesian optimization with a tree-structured Parzen estimator[25,26] All analyses were performed in Python.

Predictive model performance

We report summary statistics of model performance in predicting the composite outcome between 4 and 24 hours of hospital arrival. We used bootstrapped logistic regression with ten-fold cross validation to generate receiver operating characteristic and precision-recall benchmarks for the Elixhauser, qSOFA, CURB-65, and qCSI models and used bootstrapped gradient boosting with ten-fold cross validation to create the same metrics for the CSI model. Where necessary, data were imputed using median values of bootstraps. For significance testing, we applied Welch's t-test to average differences between permutation tests of models' performance metrics.[27,28] ROC curves describe the relationship between model sensitivity and specificity as each point represents model sensitivity and specificity at a specific cutoff. The area under these curves (AU-ROC) are a common and facile metric for comparing models to one another. Precision recall curves are an alternate metric that shows the relationship between precision (inversely related to the false positive rate), and recall (inversely related to false negative rate). AU-ROC is presented for the qCSI and CSI models as applied to the external validation cohort.

Web interface design

The qCSI was made publicly available as a web calculator at covidseverityindex.org. Nodejs, Vue, and Vuetify were used for the website frontend, while the backend was built on python using Flask.

Patient and public involvement

This was a retrospective observational cohort study and no patients were directly involved in the study design, setting the research questions, or the outcome measures. No patients were asked to advise on interpretation or presentation of results. This study was approved by our local institutional review board (IRB# 2000027747).

Results

Between March 1, 2020 and April 27, 2020, there were a total of 1,792 admissions for COVID-19 patients. Of these, 620 patients (35%) were excluded by meeting critical respiratory illness endpoints within 4 hours of presentation or by age criteria. Of the included patients, 144 (12.3%) had respiratory decompensation within the first 24 hours of hospitalization including: 101 (8.6%) requiring >10 liters/minute oxygen flow, 112 (9.6%) on a high flow device (e.g., non-rebreather, high-flow nasal cannula), 4 (0.3%) on non-invasive ventilation, 10 (0.8%) with invasive ventilation, and 1 (0.01%) death. 59 (5%) of patients were admitted to the ICU with the 4 to 24 hour time period. Population characteristics including demographics and comorbidities for the study are shown in Table 1. Study patient flow is shown in Figure 1.

Table 1: Characteristics of COVID-19 positive admitted patients stratified by primary outcome

Variable	Category	Early Respiratory Decompensation	
		Negative n=1028	Positive n=144
Age	18-44	105 (10.2)	19 (13.2)
	45-64	340 (33.1)	60 (41.7)
	> 65	583 (56.7)	65 (45.1)
Sex	Female	506 (49.2)	61 (42.4)
	Male	522 (50.8)	83 (57.6)
Race	Black or African American	260 (25.3)	40 (27.8)
	White or Caucasian	517 (50.3)	63 (43.8)
	Other	251 (24.4)	41 (28.5)
Ethnicity	Hispanic or Latino	233 (22.7)	44 (30.6)
	Non-Hispanic	776 (75.5)	97 (67.4)
	Unknown	19 (1.8)	3 (2.1)
Smoking Status	Former Smoker	340 (33.1)	45 (31.2)
	Never Smoker	503 (48.9)	66 (45.8)
	Unknown	185 (18.0)	33 (22.9)
Financial Class	Commercial	118 (11.5)	21 (14.6)
	Medicaid	136 (13.2)	23 (16.0)
	Medicare	590 (57.4)	68 (47.2)
	Other	92 (8.9)	19 (13.2)
	Self-pay	92 (8.9)	13 (9.0)
Comorbidities	Hypothyroidism	186 (18.1)	22 (15.3)
	Metastatic disease	66 (6.4)	9 (6.2)
	Other neurologic disorders	320 (31.1)	36 (25.0)
	Renal disease	205 (19.9)	30 (20.8)
	Congestive heart failure	203 (19.7)	20 (13.9)
	Depression	260 (25.3)	31 (21.5)
	Chronic pulmonary disease	282 (27.4)	32 (22.2)
	Hypertension with complications	264 (25.7)	36 (25.0)
	Valvular disease	235 (22.9)	21 (14.6)
	Anemia from blood loss	68 (6.6)	7 (4.9)
	Peripheral vascular disease	220 (21.4)	31 (21.5)
	Fluid and electrolyte disorders	378 (36.8)	47 (32.6)
	Psychoses	126 (12.3)	16 (11.1)
	Rheumatoid arthritis/collagen vascular	74 (7.2)	11 (7.6)
	Diabetes with chronic complications	263 (25.6)	37 (25.7)
	Weight loss	158 (15.4)	18 (12.5)
	Deficiency anemias	315 (30.6)	48 (33.3)
	Obesity	261 (25.4)	40 (27.8)
	Diabetes without chronic complications	93 (9.0)	19 (13.2)
	Alcohol abuse	63 (6.1)	7 (4.9)
	Drug abuse	51 (5.0)	12 (8.3)
	Liver disease	97 (9.4)	15 (10.4)
	Coagulation deficiency	98 (9.5)	10 (6.9)
	Hypertension	311 (30.3)	47 (32.6)
	Solid tumor without metastasis	96 (9.3)	10 (6.9)
	Paralysis	71 (6.9)	9 (6.2)
	Chronic peptic ulcer disease	52 (5.1)	6 (4.2)
	Pulmonary circulation disorders	64 (6.2)	6 (4.2)
	Lymphoma	12 (1.2)	—
	AIDS	16 (1.6)	2 (1.4)

Table 2: qCSI and CSI model variables. † Pulse oximetry represents the lowest value recorded during the first four hours of the patient encounter.

qCSI variable		Points	Additional CSI variables
Respiratory Rate			Aspartate transaminase
	≤ 22	0	Alanine transaminase
	23-28	1	Ferritin
	> 28	2	Procalcitonin
Pulse Oximetry†			Chloride
	$> 92\%$	0	C-reactive protein
	89-92%	2	Glucose
	$\leq 88\%$	5	Urea Nitrogen
O2 Flow Rate			White blood cell count
	≤ 2	0	Age
	3-4	4	
	5-6	5	

Identification of predictive factors for critical illness

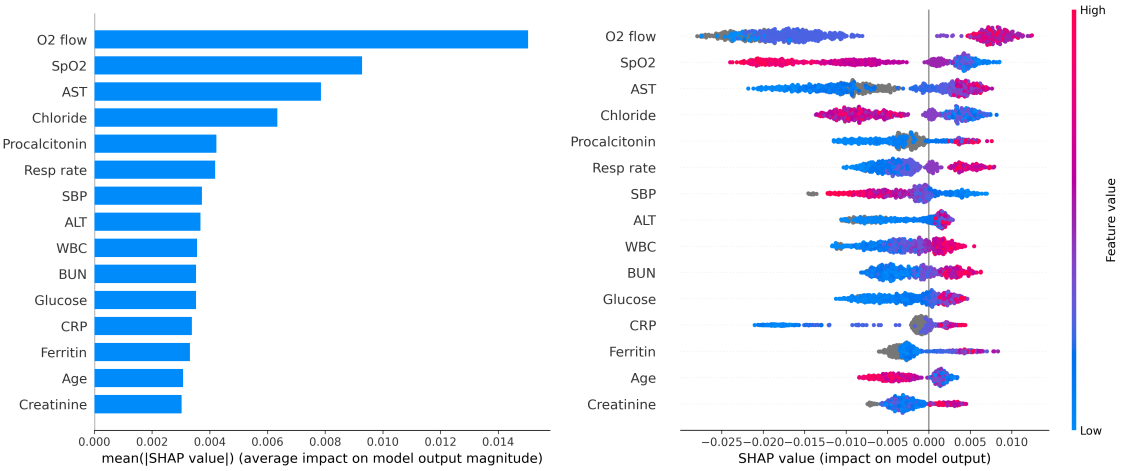
Our full dataset included 713 patient variables available during the first four hours of the patient encounters. Notably, these included demographics, vital signs, laboratory values, comorbidities, chief complaints, outpatient medications, tobacco use histories, and CXR. Radiologist evaluated CXRs was classified into three categories with strong inter-rater agreement (κ = 0.81). Our ensemble approach revealed three clinical variables as consistently imporatnt across the variable selection models: nasal cannula requirement, minimum recorded pulse oximetry, and respiratory rate.

qCSI and CSI variable weights

We divide each of these three clinical variables into value ranges using clinical experience and used logistic regression to create weights for the qCSI scoring system (2). Normal physiology was used as the baseline category, and the logistic regression odds ratios were offset to assign normal clinical parameters zero points in the qCSI.

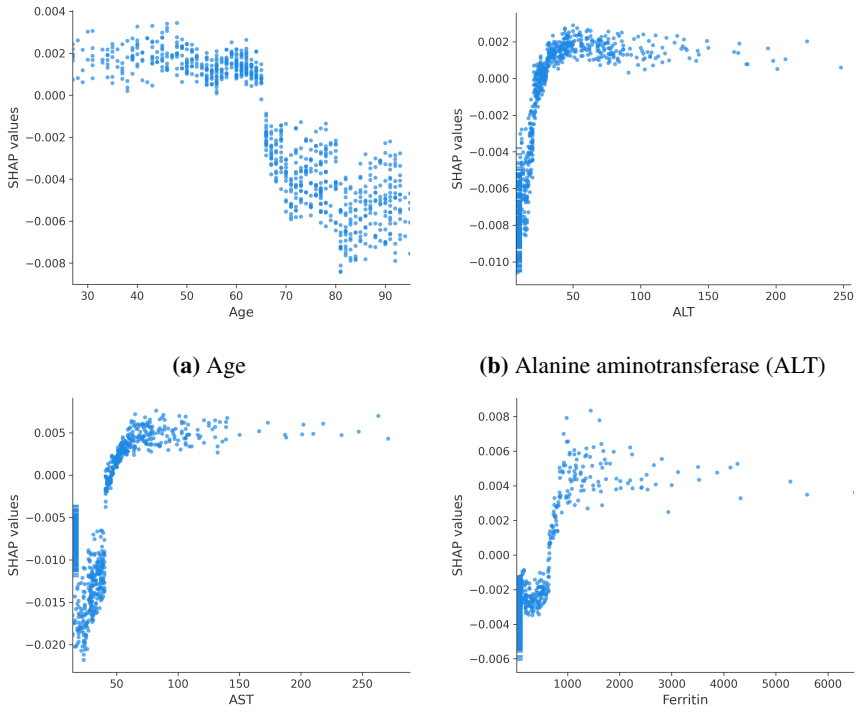
We identified an additional twelve features from the predictive factor analysis for use in a machine learning model (CSI) with gradient boosting (2). We used SHAP methods to understand the importance of various clinical variables in the CSI (Figure 2).[24,29–31] SHAP values are an extension of the game-theoretic Shapley values that seek to describe variable impacts on model output, as defined as the contribution of a specific variable to the prediction itself.[29] The key advantage of the related SHAP values is that they add interpretability to complex models like gradient boosting, which otherwise provide opaque outputs. SHAP values are dimensionless and represent the marginal contribution a variable makes on a single prediction. In the case of our gradient boosting CSI model, we employ an isotonic regression step for model calibration, so the SHAP values provide a relative weighting of contributions.[32]

Calculating the average absolute value over SHAP values suggests the most important variables in a given model - for the CSI these were flow rate by nasal cannula, followed by lowest documented pulse oximetry, and AST (fig: featureimportance). Consistent with prior studies, we also observed utility to inflammatory markers, ferritin, procalcitonin, and CRP. We then explored how ranges of individual feature values affected model output 2). For example, low oxygen flow rates (blue) are protective as indicated by negative SHAP values, as are high pulse oximetry values (red). To better investigate clinical variable effects on predicted patient risk, we generated individual variable SHAP value plots (3). Age displayed a nearly binary risk distribution with an inflection point between 60 and 70 years of age. Younger patients displayed a higher risk of 24 hour critical illness than did older patients. We also observed that elevated AST, ALT, and ferritin were associated with elevated model risk, but the SHAP values reached their asymptotes well before the maximum value for each of these features. AST and ALT SHAP values reached their maximum within normal or slightly elevated ranges for these laboratory tests. The inflection point in risk attributable to ferritin levels, however, was close to 1000 ng/mL, above institutional normal range for this test (30-400 ng/mL).



(a) SHAP Value Importance Plot **(b)** SHAP Value Beeswarm Plot

Figure 2: SHAP variable importance and beeswarm plots. (A) Mean absolute SHAP values suggest a rank order for variable importance in the CSI. (B) Each point corresponds to an individual person in the study. The points position on the x-axis shows the impact that feature has on the model's prediction for a given patient. Color corresponds to relative variable value.



(a) Age **(b)** Alanine aminotransferase (ALT)

(c) Aspartate aminotransferase (AST) **(d)** Ferritin

Figure 3: SHAP plots for selected variables

qCSI and CSI performance

qSOFA (AU-ROC, 95% CI; 0.76 [0.69-0.85]) was better than either Elixhauser (0.70 [0.62-0.80]) or CURB-65 (0.66 [0.58-0.77]) in predicting the composite endpoint (Table 3). After statistical testing with bootstrapping, the qCSI

Table 3: Performance Characteristics for CSI, qCSI, and comparison models on cross validation

Model	AU-ROC	Accuracy	Sensitivity	Specificity	AU-PRC	Brier score	F1	Average Precision
CURB-65	0.66 (0.58,0.78)	0.79 (0.56,0.94)	0.67 (0.29,1.00)	0.62 (0.27,0.93)	0.26 (0.09,0.44)	0.10 (0.06,0.15)	0.20 (0.00,0.36)	0.20 (0.10,0.33)
qSOFA	0.76 (0.69,0.86)	**0.88 (0.82,0.95)**	0.79 (0.62,1.00)	0.70 (0.60,0.80)	0.35 (0.09,0.62)	0.09 (0.05,0.14)	0.21 (0.00,0.46)	0.26 (0.13,0.42)
Elixhauser	0.70 (0.62,0.80)	0.71 (0.40,0.86)	0.73 (0.47,1.00)	0.67 (0.33, 0.88)	0.20 (0.09, 0.36)	0.10 (0.06, 0.15)	0.30 (0.15,0.43)	0.22 (0.11, 0.36)
qCSI	0.90 (0.85,0.96)	0.84 (0.72,0.94)	0.90 (0.70,1.00)	0.79 (0.59,0.94)	0.54 (0.27,0.76)	**0.07 (0.04,0.11)**	0.49 (0.30,0.67)	0.52 (0.30,0.72)
CSI	**0.91 (0.86,0.97)**	0.83 (0.70,0.94)	**0.94 (0.77,1.00)**	**0.82 (0.67,0.95)**	**0.56 (0.25,0.80)**	0.25 (0.25,0.28)	**0.51 (0.29,0.70)**	**0.58 (0.31,0.81)**

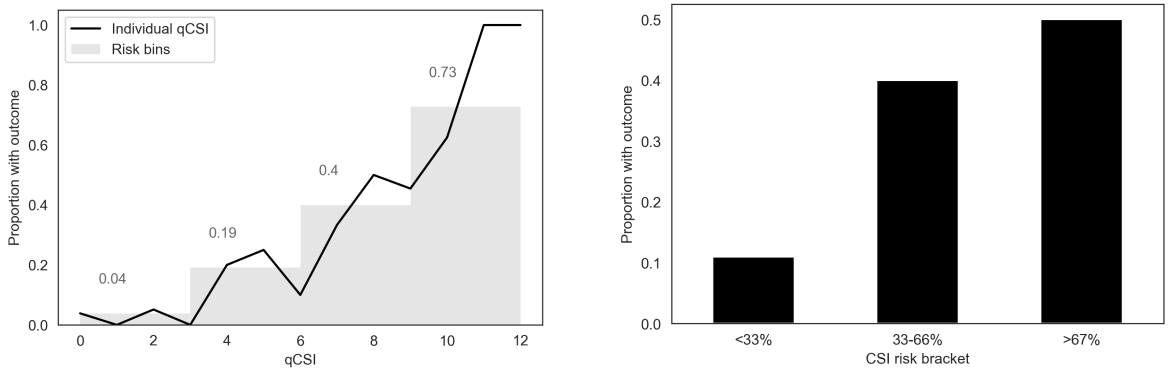
(0.89 [0.84, 0.95]) and CSI (0.92 [0.86, 0.97]) models outperformed the comparator models with the CSI best by the AU-ROC metric overall ($p < 0.05$ for all comparisons).

External validation

We then tested the predictive performance of qCSI and CSI on the external validation cohort in order to test their generalizability, finding AU-ROC of 0.82 and 0.76, respectively. We then tested the calibration of the qCSI score by assigning all patients in the external cohort a qCSI score and comparing these scores to their known outcomes (Figure 4A).[33] The calibration of the CSI was also tested on this external validation cohort (Figure 4B). These calibration curves suggest that outcome rates increased with qCSI and CSI scores.

Web application

The qCSI is available at covidseverityindex.org. The qCSI calculator includes selection boxes for each of the three variables which are summed to generate a score and prediction as estimated using the external validation cohort.



(a) Quick Covid Severity Index (qCSI)

(b) Covid Severity Index (CSI)

Figure 4: Calibration of qCSI and CSI on external validation dataset

Discussion

Consistent with clinical observations, we noted a significant rate of progression to critical respiratory illness within the first 24 hours of hospitalization in COVID-19 patients. We used six parallel approaches to identify a subset of variables for the final qCSI and CSI models. The qCSI ultimately requires only three variables, all of which are accessible at the bedside. Using this model and the calibration results on the external cohort, we proposed that a qCSI score of 3 or less be considered low-likelihood for 24 hour respiratory critical illness. We note that few patients in the validation cohort had qCSI of 3 (SpO2 of 89-92% and respiratory rates of 23-28 without any oxygen requirement) - these patients may be found to have higher risk in future studies.

While statistically significant, the modest increases in performance of the CSI as compared to the qCSI suggest that the more parsimonious qCSI is likely preferable for rapid implementation. Comparison between qCSI and CSI on the external validation cohort offers a snapshot of potential generalizability, but further studies will be required. The

CSI, however, offers opportunities to perform further analysis of potential COVID-19 prognostic factors. In alignment with current hypotheses about COVID-19 severity, we note that multiple variable selection techniques identified inflammatory markers including CRP and ferritin as potentially important predictors. More striking however, was the importance of aspartate (AST) and alanine aminotransferase (ALT) in CSI predictions as calculated with SHAP values.[34,35] Lower age had higher SHAP values, suggesting potential bias in the admitted patient cohort - young, admitted patients may be more ill than older admitted COVID-19 patients. Interestingly, the transition point where the SHAP value analysis identified model risk associated with liver chemistries was at the high end of normal, consistent with previous observations that noted that normal to mild liver dysfunction among COVID-19 patients. We hypothesize that the asymptotic quality of the investigated variables with respect to CSI risk contributions reflects our moderate study size and we expect that scaling CSI training to larger cohorts will further elucidate the impacts of more extreme values on risk. While our dataset included host risk factors including smoking history, obesity, and BMI, these did not appear to play a prominent role in predicting deterioration. Here, we recognize two important considerations: first, that predictive factors may not be mechanistic or causative factors in disease, and second that these factors may be related to disease severity without providing predictive value for 24 hour decompensation.

We include CXRs for 1,170 visits in this cohort. CXR are of significant clinical interest as previous studies have shown high rates of ground glass opacity and consolidation.[36] Chest CT may have superior utility for COVID-19 investigation, is not being widely performed at our institutions as part of risk stratification or prognostic evaluation.[37] CXR reports were classified based on containing bilateral, unilateral, or no opacities or consolidations. We found high inter-rater agreement in this coding, but CXR were not consistently identified by our variable selection models. Further studies using natural language processing of radiology reports or direct analysis of CXR with tools like convolutional neural networks will provide more evidence regarding utility of these studies in COVID-19 prognostication.[38] Furthermore, we do not consider other applications of CXR including the identification of other pulmonary findings like diagnosis of bacterial pneumonia.

The Elixhauser comorbidity mortality score, qSOFA, and CURB-65 baseline models provided the opportunity to test well-known risk stratification and prognostication tools with a COVID-19 cohort. These tools were selected, in part, for their familiarity within the medical community, and because each has been proposed as having potential utility within the COVID-19 epidemic. We note the relatively limited predictive performance of these metrics, while simultaneously recognizing that none were designed to address the clinical question addressed here. In particular, the CURB-65 pneumonia severity score may still have utility in determining patient disposition with respect to discharge or hospitalization.

Future studies will be required to expand on this work in a number of ways. First, prospective, multi-site validation is required for the qCSI. The CSI may lend itself to a "living" model framework where the addition of new features, weights, and outcomes will improve its predictive capability.[8,39] We hypothesize that the CSI will continue to improve as compared to the qCSI as more patient observations are included. Second, we expect related models to be extended to patient admission decisions as well as continuous hospital monitoring.[40–42] The qCSI does not separate patients without any nasal cannula requirement from those with even a minimal oxygen requirement. We expect that future models for safe discharge of COVID-19 patients will more strongly weigh even low oxygen requirements as local practice patterns may likely necessitate admission of any patient on exogenous oxygen.

Patient prognosis has important ramifications in terms of resource utilization, hospital placement, and patient shared decision-making. We additionally note the role of respiratory parameters in selecting patients for therapeutic interventions. An early proof-of-concept study for the viral RNA polymerase inhibitor Remdesivir, which has *in vitro* activity against SARS-CoV-2, included patients with pulse oximetry of $\leq 94\%$ on ambient air or who had any oxygen requirement.[43] There is a large ongoing clinical trial that uses similar inclusion criteria (`ClinicalTrials.gov` Identifier: NCT04292899). A 237 patient Chinese trial of the same drug was stopped early after no further eligible patients were available for enrollment.[44] This study included patients with confirmed COVID-19 infection by RT-PCR, pneumonia on imaging, oxygen saturation of $\leq 94\%$ on ambient air, or a partial pressure to fractional inspired oxygen ratio of 300 mm Hg or less. Improved pragmatic, prognostic tools like the qCSI may offer a route to expanded inclusion criteria for ongoing trials or for early identification of patients who might potentially benefit from therapeutics.

Limitations

The data in this study were observational data provided from a single health system and so may not be generalizable based on local testing and admissions practices. Our data were extracted from an electronic health record, which is associated with known limitations including propagation of old or incomplete data. Similarly, there are important markers of oxygenation which were out of the scope of our study, including alveolar-arterial gradients.

Retrospective observational studies lack control of variables so prospective studies will be required to assess validity of the presented models and the specificity of the features we identify as important to COVID-19 progression. Assumptions were made in data processing where noted in the methods, which introduce biases into our results. Chest x-ray interpretation was done manually using radiology reports, but without reviewing the radiography, which introduces subjectivity as reflected in the inter-rater agreement metric. Most significant, however, is that management of COVID-19 is evolving, so it may be possible that future clinical decisions, like when to intubate patients, may not match those standards used in the reported clinical settings.

Conclusions

The qCSI robustly predicts clinical respiratory decompensation in COVID-19 patients using pulse oximetry, respiratory rate, and nasal cannula flow rate. The CSI, a gradient boosting machine learning model, modestly improves on the qCSI and highlights the predictive performance of a number of variables including liver chemistries and inflammatory markers. Prospective, multi-site validation will be required to better assess the generalizability of these models. The qCSI is available at covidseverityindex.org.

Funding: FPW acknowledges R01DK113191 and P30DK079310.

Conflicts of interest: WLS was an investigator for a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; is a technical consultant to HugoHealth, a personal health information platform; co-founder of Refactor Health, an AI-augmented data mapping platform for healthcare; and is a consultant for Interpace Diagnostics Group, a molecular diagnostics company.

References

1. Guan Wj, Ni Zy, Hu Y, Liang Wh, Ou Cq, He Jx, et al. Clinical characteristics of coronavirus disease 2019 in China. New England Journal of Medicine. 2020;.

2. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet. 2020;.

3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497–506.

4. Bhatraju PK, Ghassemieh BJ, Nichols M, Kim R, Jerome KR, Nalla AK, et al. Covid-19 in critically ill patients in the Seattle region—case series. New England Journal of Medicine. 2020;.

5. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, et al. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy Region, Italy. JAMA. 2020;.

6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. 2020 04;Available from: https://doi.org/10.1001/jama.2020.6775.

7. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell LF, Chernyak Y, et al. Factors associated with hospitalization and critical illness among 4,103 patients with COVID-19 disease in New York City. medRxiv. 2020;Available from: https://www.medrxiv.org/content/early/2020/04/11/2020.04.08.20057794.

8. Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj. 2020;369.

9. Chan PS, Jain R, Nallmothu BK, Berg RA, Sasson C. Rapid response teams: a systematic review and meta-analysis. Archives of internal medicine. 2010;170(1):18–26.

10. Badawi O, Liu X, Berman I, Amelung PJ, Doerfler M, Chandra S. Impact of COVID-19 pandemic on severity of illness and resources required during intensive care in the greater New York City area. medRxiv. 2020;Available from: https://www.medrxiv.org/content/early/2020/04/14/2020.04.08.20058180.

11. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical care. 2009;p. 626–633.

12. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016;315(8):801–810.

13. Ferreira M, Blin T, Collercandy N, Szychowiak P, Dequin PF, Jouan Y, et al. Critically ill SARS-CoV-2-infected patients are not stratified as sepsis by the qSOFA. Annals of Intensive Care. 2020;10(1):1–3.

14. Lim WS, Van der Eerden M, Laing R, Boersma W, Karalus N, Town G, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax. 2003;58(5):377–382.

15. Ilg A, Moskowitz A, Konanki V, Patel PV, Chase M, Grossestreuer AV, et al. Performance of the CURB-65 score in predicting critical care interventions in patients admitted with community-acquired pneumonia. Annals of emergency medicine. 2019;74(1):60–68.

16. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015;162(1):W1–W73.

17. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Journal of clinical epidemiology. 2008;61(4):344–349.

18. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. Journal of Medical Internet Research. 2019 Apr;21(4):e13043.

19. DataBank F. First DataBank Enhanced therapeutic classification system (ETC). First Databank; 2020. http://www.firstdatabank.com/Products/therapeutic-classification-system-nddf.aspx.

20. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care. 1998;p. 8–27.

21. for Healthcare Research A, Quality. HCUP Tools and Software. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality, Rockville, MD; 2020. www.hcup-us.ahrq.gov/tools_software.jsp.

22. Cohen SB, Ruppin E, Dror G. Feature Selection Based on the Shapley Value. In: IJCAI. vol. 5; 2005. p. 665–670.

23. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157–1182.

24. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):2522–5839.

25. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–794.

26. Bergstra J, Yamins D, Cox DD. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML'13. JMLR.org; 2013. p. I–115–I–123.

27. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability. Boca Raton, Florida, USA: Chapman & Hall/CRC; 1993.

28. Janssen A, Pauls T. How do bootstrap and permutation tests work? Annals of Statistics. 2003 06;31(3):768–806.

29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems; 2017. p. 4765–4774.

30. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering. 2018;2(10):749.

31. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nature Medicine. 2020;26(1):71–76.

32. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning; 2005. p. 625–632.

33. Backus B, Six A, Kelder J, Bosschaert M, Mast E, Mosterd A, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. International journal of cardiology. 2013;168(3):2153–2158.

34. Zhang C, Shi L, Wang FS. Liver injury in COVID-19: management and challenges. The Lancet Gastroenterology & Hepatology. 2020;.

35. Cai Q, Huang D, Yu H, Zhu Z, Xia Z, Su Y, et al. Characteristics of Liver Tests in COVID-19 Patients. Journal of Hepatology. 2020;.

36. Wong HYF, Lam HYS, Fong AHT, Leung ST, Chin TWY, Lo CSY, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology. 2020;p. 201160.

37. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology. 2020;295(1):202–207.

38. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. CoRR. 2017;abs/1711.05225. Available from: http://arxiv.org/abs/1711.05225.

39. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS medicine. 2014;11(2).

40. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Science translational medicine. 2015;7(299):299ra122–299ra122.

41. Simonov M, Ugwuowo U, Moreira E, Yamamoto Y, Biswas A, Martin M, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: A descriptive modeling study. PLoS medicine. 2019;16(7).

42. Tomašev 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(7767):116–119.

43. Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, et al. Compassionate use of remdesivir for patients with severe Covid-19. New England Journal of Medicine. 2020;.

44. Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. The Lancet. 2020;.