
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

# An Interpretable Chest CT Deep Learning Algorithm for Quantification of COVID-19 Lung Disease and Prediction of Inpatient Morbidity and Mortality

Jordan H. Chamberlin, BS, Gilberto Aquino, MD, Uwe Joseph Schoepf, MD, Sophia Nance, BS, Franco Godoy, BS, Landin Carson, BS, Vincent M. Giovagnoli, BS, Callum E. Gill, MS, Liam J. McGill, BS, Jim O'Doherty, PhD, Tilman Emrich, MD, Jeremy R. Burt, MD, Dhiraj Baruah, MD, Akos Varga-Szemes, MD, PhD, Ismail M. Kabakus, MD, PhD

**Rationale and Objectives:** The burden of coronavirus disease 2019 (COVID-19) airspace opacities is time consuming and challenging to quantify on computed tomography. The purpose of this study was to evaluate the ability of a deep convolutional neural network (dCNN) to predict inpatient outcomes associated with COVID-19 pneumonia.

**Materials and Methods:** A previously trained dCNN was tested on an external validation cohort of 241 patients who presented to the emergency department and received a chest computed tomography scan, 93 with COVID-19 and 168 without. Airspace opacity scoring systems were defined by the extent of airspace opacity in each lobe, totaled across the entire lungs. Expert and dCNN scores were concurrently evaluated for interobserver agreement, while both dCNN identified airspace opacity scoring and raw opacity values were used in the prediction of COVID-19 diagnosis and inpatient outcomes.

**Results:** Interobserver agreement for airspace opacity scoring was 0.892 (95% CI 0.834-0.930). Probability of each outcome behaved as a logistic function of the opacity scoring (25% intensive care unit admission at score of 13/25, 25% intubation at 17/25, and 25% mortality at 20/25). Length of hospitalization, intensive care unit stay, and intubation were associated with larger airspace opacity score ($p$ = 0.032, 0.039, 0.036, respectively).

**Conclusion:** The tested dCNN was highly predictive of inpatient outcomes, performs at a near expert level, and provides added value for clinicians in terms of prognostication and disease severity.

**Key Words:** COVID-19; Artificial Intelligence; Thoracic Radiology; Critical Care; Pulmonology.



**Abbreviations:** **AI** Artificial intelligence, **AIC** Akaike information criterion, **ANOVA** Analysis of variance, **AUC** Area under curve, **BMI** Body mass index, **CI** Confidence interval, **COVID-19** Coronavirus disease 2019, **CT** Computed tomography, **dCNN** deep convolutional neural network, **ED** Emergency department, **HU** Hounsfield units, **ICC** Intraclass correlation coefficient, **ICU** Intensive care unit, **PCR** polymerase chain reaction, **SARS-CoV-2** Severe acute respiratory syndrome coronavirus 2

## INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has created a unique challenge for medical personnel worldwide by becoming quickly pervasive. Many studies have identified the signs found and usefulness of chest computed tomography (CT) imaging (or even abdominopelvic lung base analysis) for triage of these patients with potential COVID-19 pneumonia, particularly to identify diagnostic and prognostic factors (1–4). Therefore, the use of artificial intelligence (AI) deep learning models to

**Acad Radiol 2022; 29:1178–1188**

From the Division of Cardiovascular Imaging, Department of Radiology and Radiological Science, Medical University of South Carolina, 25 Courtenay Drive Room 2221 ART, Charleston, SC 29425 (J.H.C., G.A., U.J.S., S.N., F.G., L.C., V.M.G., C.E.G., L.J.M., T.E., J.R.B., D.B., A.V.-S., I.M.K.); Siemens Healthineers, Princeton, New Jersey (J.O.). Received December 29, 2021; revised March 17, 2022; accepted March 24, 2022. **Address correspondence to:** U.J.S. e-mail: schoepf@musc.edu


https://doi.org/10.1016/j.acra.2022.03.023

prognosticate from CT images has been identified from the beginning of the pandemic as a potential way to expedite the triage process, improve prognostication, and guideline utilization of resources (1, 5). The use of AI to prognosticate clinical course of COVID-19 pneumonia patients from subjective imaging features is challenging. One solution is the use of scoring systems, such as severity scoring, as standardization and efficiency are increased by protocol, resulting in higher-quality, evidence-based decision making by clinicians. However, manual segment severity scoring is a time-consuming task which is not currently standard of care. Thus, utilizing AI severity scoring may be helpful in meeting the challenge of practical, reproducible triage of COVID-19 patients by identifying patients at high risk for morbidity and mortality (6).

While there is a relative paucity of studies utilizing severity scoring during the task of COVID-19 CT image interpretation, several studies have demonstrated the efficacy of scoring images with severity scoring methods (6-9). Lessman, et al. reported moderate agreement for score determination by AI methods when in comparison to expert radiologists' interpretation; with high area under curve (AUC), sensitivity, and specificity (internal set: 0.95, 85.7%, and 89.8% and external set: 0.88, 82.0%, and 80.5%, respectively) (6). Goncharov, et al. demonstrated an AUC of 0.95 and severity model correlation 0.98 for the identification of COVID-19 pneumonia patients (7). Lassau, et al. calculated severity scores based on clinical factors and then recalculated the scored based on the combination of clinical factors and imaging interpretation by AI. The AI-assisted method of score calculation outperformed the previously determined score in terms of prognostic ability (8). Finally, Mader et al. determined that severity scoring differentially predicts patients with severe disease from non-severe (10).

However, methods and image sources vary between studies and may be prone to bias and overfitting from use of identical or poorly annotated images from publicly available datasets. Furthermore, few seldom test their methods against a real-world contiguous patient cohort with well-defined outcomes (1). Therefore, the purpose of this study is to analyze the efficacy of a novel deep learning model in determining prognostic value of an AI severity scoring algorithm.

## METHODS

### Study Population, Clinical Information, and Imaging Data Acquisition

The protocol of this retrospective study was approved by the local Institutional Review Board and the need for informed consent was waived. Patient data was collected and anonymized in compliance with HIPAA and institutional protocols to protect patient privacy. A total of 241 patients were enrolled in this study, 93 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) polymerase chain reaction (PCR) positive and 158 SARS-CoV-2 PCR negative, who underwent a chest CT with or without contrast from March 2020 to February 2021. Data collected included demographics, clinical comorbidities, and outcome variables which included hospitalization, intensive care unit (ICU) admission, intubation, and mortality. A preliminary patient list was collected through billing code search using COVID-19 testing and chest CT identifiers. Data collection was performed by chart review and compiled in a de-identified encrypted document. Imaging data from chest CT scans with 1 mm slice thickness including non-contrast and iodinated contrast enhanced studies (mAs and kVp selected according to patients' body mass index) were acquired from Somatom Force and Naeotom Alpha CT scanners (Siemens Healthineers, Forcheim, Germany). Archived data was then exported from the picture archiving and communication system and uploaded to the AI interface (AI-RAD companion, Siemens Healthineeers) where the algorithm was executed and the results recorded.

### Study Design

A single-institution retrospective case-control study was performed. Inclusion criteria included patients >18 years old who presented to the emergency department, received both a COVID-19 test and a chest CT within 14 days, and had sufficient same-institution follow-up for outcomes analysis (1 month post discharge from emergency department (ED) or inpatient hospitalization). Controls were selected based on an eligible CT scan with a negative SARS-CoV-2 PCR in the stated timeframe. These controls were neither age nor sex matched. Exclusion criteria included prior pulmonary surgical history, viral pneumonia other than COVID-19, and excessive artifact on chest CT.

The gold standard used was an expert-derived airspace opacity score. Three cardiothoracic trained radiologists comprised the expert determination of airspace opacities as given by Bernheim et al. (11) For each lobe, the disease extent was judged to be one of the following categories: (0) the lobe is not affected; (1) 1%-25%; (2) 25%-50%; (3) 50%-75%; and (4) 75%-100%. The scores for each of the five lobes were summed to calculate the total severity score, resulting in a total score range from 0 to 20. A 0 indicates that none of the lobes are involved and 20 indicates that all five lobes are severely affected.

The primary endpoints were interobserver agreement between AI and the radiologists for the determination of COVID-19 extent as well as the predictive capability of airspace opacity scoring and other AI measurements for the diagnosis of COVID-19 pneumonia.

### Convolutional Neural Network Architecture and Outputs

The deep convolutional neural network (dCNN) algorithm has been previously described in Chaganti et al. (12) Briefly, the original dCNN was trained on 901 chest CT scans (431 COVID-19, 174 viral pneumonia, and 296 with interstitial

lung disease) with a validation cohort of 200 patients (100 COVID-19 and 100 control). The general architecture utilized a preprocessing step with deep-image-to-image lung segmentation using the carina as a landmark with alignment, then a DenseUNet architecture for feature (ground glass opacity, etc.) extraction, subsequently followed by segmentation and global classification. Please see appendix E1 in Chaganti et al. for a detailed description of the neural network architecture, training, and measures such as loss function.

### Statistical Analysis

A power calculation optimized for outcomes assuming at least a 10% prevalence of each event required >150 patients for a standardized power of 0.8. Post-hoc, 241 patients conferred a power of 0.965 for simple logistic regression analysis (**Fig S1**). Aggregate demographics and clinical risk factors analysis was performed using SARS-CoV-2 PCR positivity as the stratifying variable. Continuous variables were assessed for normality and reported as medians plus interquartile ranges. Categorical variables were reported with count and frequency as percent.

Primarily, interobserver agreement for quantitative scoring was assessed using intraclass correlation coefficients (ICC) with 2-way mixed effects, single rater (k), and absolute agreement. Adjusted linear model $R^2$ and *p*-values were also reported for assessment of linearity of results. Cohen's kappa was reported with confidence interval as a secondary measure of categorical agreement. For categorical agreement, any airspace opacity was counted as a positive result, and no airspace opacities were defined as the only negative result and only used in the context of COVID-19 positive patients to focus on specific performance on COVID-19 patients. Diagnostic parameters were reported using confidence intervals constructed using the Clopper-Pearson method.

Multivariate modelling for COVID-19 diagnosis was performed using multiple logistic regression. Briefly, backwards stepwise logistic regression was performed on all AI generated measurements until all retained model elements were significant ($p < 0.05$) in the model. The model with the lowest Akaike information criterion was selected among the models with significant elements. Multivariate modelling of outcomes was performed using the variables deemed diagnostic for COVID-19 in the previous analysis. Optimal airspace opacity score cutoffs were empirically selected using a bootstrapping approach with 200 repetitions of 1:1 COVID-19 positive/negative stratification sampling were used by maximization of the bootstrapped accuracy metric. **Figure S2** demonstrates the empiric selection process. Time-derived outcome variables were analyzed by binning into quintiles to improve reader interpretability. Differences between each quintile were assessed using one-way ANOVA. Means and standard errors were reported for continuous variables. All statistical analysis was performed in R v 3.6.3.

## RESULTS

In this study 93 patients (38.5%) were positive for SARS-CoV-2. The median age of those with and without COVID-19 was 59 (IQR 45-71) and 62 (IQR 47-69), respectively. A greater proportion of those with COVID-19 were male in comparison to controls (61.5% vs 51.9). The median time between nasopharyngeal swab and imaging was 3 days for SARS-CoV-2 positive patients and 0 days for SARS-CoV-2 negative patients. Patients positive for SARS-CoV-2 were more likely to be Black or Hispanic (57.0%, 2.3%) than SARS-CoV-2 negative patients (37.7%, 0%). In comparison to control patients, SARS-CoV-2 positive patients were more frequently smokers (93.4% vs 48.8%), more likely to have hypertension (65.9% vs 49.2%), and more likely to be diabetic (37.6% vs 23.8) **(Table 1)**.

The AI dashboard, the provided summary of the algorithm output, demonstrates highlighted airspace opacities in the axial view with the possibility to reconstruct the affected tissue in three dimensions. The results dashboard provides readers with information regarding the extent of the airspace opacities as broken down by lobe. Results include opacity scores, lung volumes, mean and standard deviation of the Hounsfield units for affected lungs, volumes of affected lung tissue and high opacity measurements **(Fig 1).**

The overall correlation between observer estimates of severity score was 0.827 (95% CI 0.751 - 0.891). The expert and AI had a high rate of agreement with ICC of 0.892 (95% CI 0.834-0.930), $p < 0.001$. The Adjusted $R^2$ for explanation of model variance was 0.69, $p < 0.001$ **(Fig 2a).** Overall, The accuracy of the dCNN was 0.828 (95% CI 0.751-0.905) and sensitivity was 0.914 (95% CI 0.830-0.965) **(Fig 2b).**

Using the measurements given in the AI dashboard, a best fit multivariate model consisting of total opacity volume ($cm^3$), high opacity volume ($cm^3$), standard deviation of opacity Hounsfield units, and total standard deviation of all Hounsfield units gives an AUC of 0.805 (95% CI 0.745-0.862) for the diagnosis of COVID-19. The AUC for individual predictors of the model range from 0.728 to 0.561. All variable coefficients were significant in the multiple logistic regression model ($p < 0.05$) **(Fig 3).** The same combination of variables predicts need for inpatient hospitalization (AUC = 0.810) and ICU admission, intubation, and mortality at AUCs ranging from 0.666 to 0.683. The trend of accuracy was highest for events earliest in each patient's time course **(Fig 4).**

In regards to threshold determination, AI Airspace opacity ≥13 was accurate (0.777 95%; CI 0.724-0.829) and specific (0.873; 95% CI 0.822-0.913) for mortality. AI airspace opacity ≤13 had a high NPV for death (0.946; 95% CI 0.915-0.977). Accuracy of AI airspace opacity ≥8 for hospitalization was 0.777 (95% CI 0.724-0.829). Accuracy of AI airspace opacity ≥ 9 for ICU admission was 0.744 (95% CI 0.680-0.799). Accuracy of AI airspace opacity ≥12 for intubation was 0.839 (95% CI 0.793-0.885) **(Table 2)**. Using the threshold values within the logistic model probabilities, there

**TABLE 1. Demographics and Clinical Comorbidities of Patients Enrolled in this Study Stratified by SARS-CoV-2 Nasopharyngeal Swab PCR Results**

| N = 241 | SARS-CoV-2 Positive (N = 93) | | SARS-CoV-2 Negative (N = 148) | |
|---|---|---|---|---|
| | Median | IQR | Median | IQR |
| Age (years) | 59 | 45-71 | 62 | 47-69 |
| BMI (kg/m$^2$) | 29.3 | 25.8-36.1 | 26.5 | 21.9-33.0 |
| Symptom days | 6 | 2-9 | 4 | 1-9 |
| PCR-Imaging Δ | 3 | 0-8 | 0 | 1-9 |
| | Count | Frequency (%) | Count | Frequency (%) |
| *Sex* | | | | |
| Female | 35 | 38.5 | 62 | 48.1 |
| Male | 56 | 61.5 | 67 | 51.9 |
| *Ethnicity* | | | | |
| Black | 49 | 57.0 | 40 | 37.7 |
| Hispanic | 2 | 2.3 | 0 | 0 |
| Other | 5 | 5.8 | 2 | 1.9 |
| White | 30 | 34.9 | 64 | 60.4 |
| Prior Structural Lung disease | 32 | 34.8 | 37 | 28.5 |
| History of Cancer | 9 | 11.9 | 46 | 33.6 |
| Smoking History | 85 | 93.4 | 44 | 48.4 |
| Hypertension | 60 | 65.9 | 64 | 49.2 |
| Diabetes | 34 | 37.6 | 31 | 23.8 |
| CHF | 16 | 17.8 | 27 | 20.8 |
| CKD | 14 | 15.4 | 19 | 14.6 |
| Autoimmune disease | 14 | 15.4 | 17 | 13.1 |
| HIV | 0 | 0 | 6 | 4.6 |

BMI, body mass index; CHF, congestive heart failure; CKD, chronic kidney disease; HIV, human immunodeficiency virus; IQR, interquartile range; PCR, polymerase chain reaction; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

is a 25% risk of respective outcomes at an airspace score of 6 (Hospitalization), 13 (ICU Admission), 16 (Intubation) and ≥20 (Mortality). Significant increases in probability of mortality does not occur until AI Airspace opacity >10. A maximum score of 20 conferred an 87.5% probability of hospitalization, 50% probability of ICU admission, 37.5% probability of intubation, and about a 25% probability of mortality. The points of most uncertainty came at scores > 15, suggesting other risk factors are increasingly important at these upper ranges **(Fig 5)**.

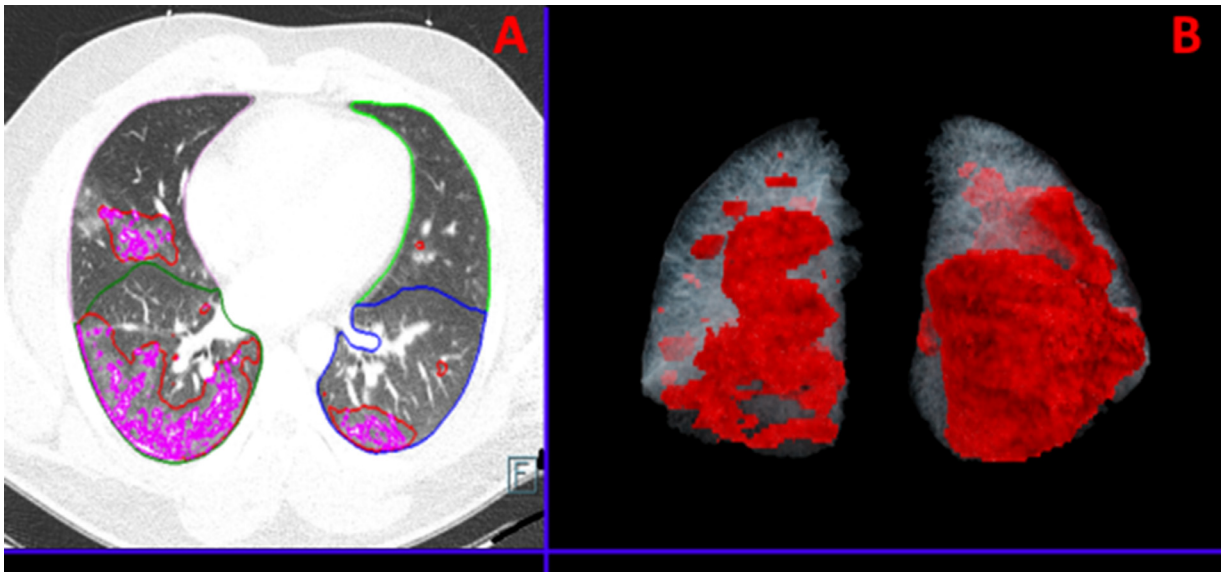
The AI airspace opacity scores predict time-to-event and inpatient durations, with the mean hospitalization duration, ICU duration, and intubation duration being associated with increased AI airspace opacity scores ($p$ = 0.032, 0.039, 0.036, respectively). The time from hospital admission to ICU admission was not significantly associated with AI airspace opacity scores ($p$ = 0.159) **(Fig 6)**.

## DISCUSSION

The purpose of this study was to test a previously trained deep convolutional neural network for diagnostic and prognostic purposes in patients with COVID-19 pneumonia as seen on chest CT. A total of 241 patients (93 COVID-19 positive) were evaluated by the dCNN in this external testing cohort design. The AI algorithm was highly accurate compared to attending radiologists with ICCs approaching human-level agreement. Several key interpretable outputs were derived including opacity volumes, parenchymal-opacity ratios, and other 2nd order statistics. When put together into a standardized scoring system, several cutoffs were identified that process in a stepwise fashion in terms of severity. Lastly, both probabilities of inpatient outcomes and time-to-events behaved as a function of the airspace opacity scoring system, establishing expected prognostic gradients that may influence patient care.

It is critical to understand the accuracy of expert observers in the diagnosis of COVID-19 pneumonia from chest CT, as the gold standard used in this study was the expert quantification of airspace severity. Baseline expert accuracy in comparison to PCR surpasses 90% for the diagnosis of COVID-19 pneumonia. The ICC for expert-AI quantitative severity scoring represented "excellent" agreement. Overall, AI accuracy for patients with COVID-19 by positive PCR was high for identifying airspace opacities related to COVID-19 lesions. However, the correlation coefficient in this external validation cohort was mildly less than the previously published training data for the neural network (12).

While the focus of this study is on AI severity scoring, multivariate modelling of AI segmented measurements has an advantage over a scoring heuristic for the diagnosis of COVID-19 pneumonia (13, 14). A multivariate model consisting of opacity volume, "high opacity" volume, and the standard deviation of both opacity Hounsfield units and total

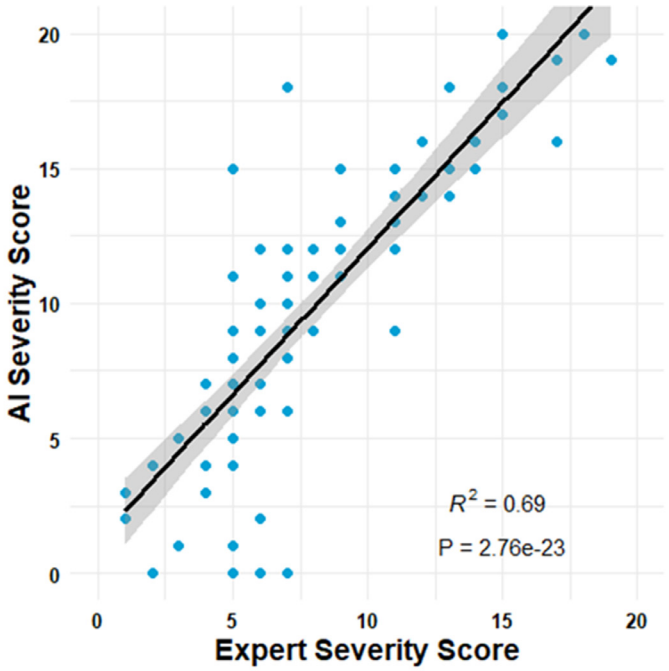



| LUNG OPACITY | Both Lungs | Right Lung | Left Lung |
|---|---|---|---|
| Opacity Score | 7 | 5 | 2 |
| Total Volume [ml] | 2879.97 | 1529.71 | 1350.27 |
| Opacity Volume [ml] | 512.54 | 387.93 | 124.61 |
| Opacity Percentage [%] | 17.80 | 25.36 | 9.23 |
| High Opacity Volume [ml] | 118.55 | 99.55 | 19.01 |
| High Opacity Percentage [%] | 4.12 | 6.51 | 1.41 |
| Mean HU Total [HU] | -596.65 | -579.94 | -615.59 |
| Mean HU of Opacity [HU] | -381.68 | -366.46 | -429.06 |
| Standard Deviation Total [HU] | 251.83 | 273.29 | 223.56 |
| Standard Deviation of Opacity [HU] | 231.90 | 235.33 | 214.06 |


| LUNG LOBE OPACITY | Right Upper | Right Middle | Right Lower | Left Upper | Left Lower |
|---|---|---|---|---|---|
| Opacity Score | 1 | 1 | 3 | 1 | 1 |
| Total Volume [ml] | 611.67 | 373.54 | 544.50 | 810.30 | 539.97 |
| Opacity Volume [ml] | 80.86 | 20.38 | 286.69 | 9.32 | 115.29 |
| Opacity Percentage [%] | 13.22 | 5.45 | 52.65 | 1.15 | 21.35 |
| High Opacity Volume [ml] | 13.06 | 3.63 | 82.86 | 0.87 | 18.14 |
| High Opacity Percentage [%] | 2.13 | 0.97 | 15.22 | 0.11 | 3.36 |
| Mean HU Total [HU] | -640.53 | -711.83 | -421.39 | -674.73 | -526.85 |
| Mean HU of Opacity [HU] | -465.77 | -449.93 | -332.51 | -476.23 | -425.25 |
| Standard Deviation Total [HU] | 235.94 | 221.88 | 268.92 | 199.61 | 228.17 |
| Standard Deviation of Opacity [HU] | 240.70 | 246.10 | 223.29 | 193.61 | 215.18 |

**Figure 1.** Artificial Intelligence dashboard for automated evaluation of chest computed tomography for COVID-19. (**a**) Axial view of lung fields with highlighted opacities segmented by neural network algorithm. (**b**) Three-dimensional Image reconstruction of lungs with rending of involved airspace opacities. (**C**) Parameters involved with diagnosis of COVID-19 by AI. AI, Artificial intelligence; COVID-19, Coronavirus disease 2019. (Color version of figure is available online.)

A




B

| Interobserver agreement between AI and Expert Amongst COVID Positive patients (N = 93) | | | |
|---|---|---|---|
| | Expert Airspace Opacity | No Airspace Opacity | |
| AI Airspace Opacity | 74 | 9 | |
| No Airspace Opacity | 7 | 3 | |
| Accuracy | 0.828 (0.751 - 0.905) | Cohen's Kappa | 0.176 (0 - 0.543) |
| Sensitivity | 0.914 (0.830 - 0.965) | PPV | 0.982 (0.825 - 0.958) |
| Specificity | 0.250 (0.055 - 0.572) | NPV | 0.300 (0.016 - 0.584) |

**Figure 2.** Interobserver agreement between expert and AI opacity scores in patients who were positive for SARS-CoV-2 by PCR. (**a**) Quantitative comparison of opacity score. (**b**) Qualitative assessment for detection of any airspace opacities. AI, Artificial intelligence; NPV, negative predictive value; PCR, polymerase chain reaction; PPV, positive predictive value; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2. (Color version of figure is available online.)

Hounsfield units provides an AUC of 0.805, greater than the sum of its parts or the opacity scoring system. Expert measurement of opacity volumes and standard deviations are not feasible, reflecting a possible advantage of using AI systems in the prediction of COVID-19 pneumonia. Indeed, some radiomic studies suggest the quantitative parenchymal involvement to be important indicators of severe outcomes (15-17). The loss of accuracy from the severity scoring system can be attributed to the trade of interpretability for accuracy in any scaled heuristic (18).

Further clinical utility can be derived from the prediction of outcomes from airspace severity scoring. Quantitative AI airspace values readily predict inpatient hospitalization with reasonable accuracy, providing immediate clinical utility from the emergency department. More advanced outcomes (ICU admission, Intubation, and mortality) had predictions which were less strong, likely related to the multifactorial risk factors for each outcome. Certainly, already verified risk factors such as age, immunosuppression, BMI, and sex contribute to the overall predictive value of the imaging factors to a large degree in late inpatient clinical outcomes.

The presence of "large" or "extensive" airspace opacities on chest imaging often evokes a negative reaction for poor prognosis among physicians caring for COVID-19 patients. However, the actual relationship of the quantitative extent of the airspace opacities and inpatient outcomes is poorly understood (18). Certainly, radiologists may be able to segment airspace opacities by hand to provide extra clinical value, but this is a time intensive and laborious process which presents difficulty in the setting of increased chest imaging volumes during the COVID-19 pandemic (19). Therefore, the introduction of an AI algorithm that would automatically segment the airspace opacities and provide a numeric, interpretable score could add value to the prognostication of COVID-19 pneumonia and change clinical management as patients progress down the COVID-19 treatment protocol (20). Several such algorithms have been proposed with ICCs above the 90th percentile, each deriving value from individual lobar involvement (21, 22). Still, a main challenge with expert-derived approaches include interobserver variation, which is partially rectified using a standardized AI approach (23).

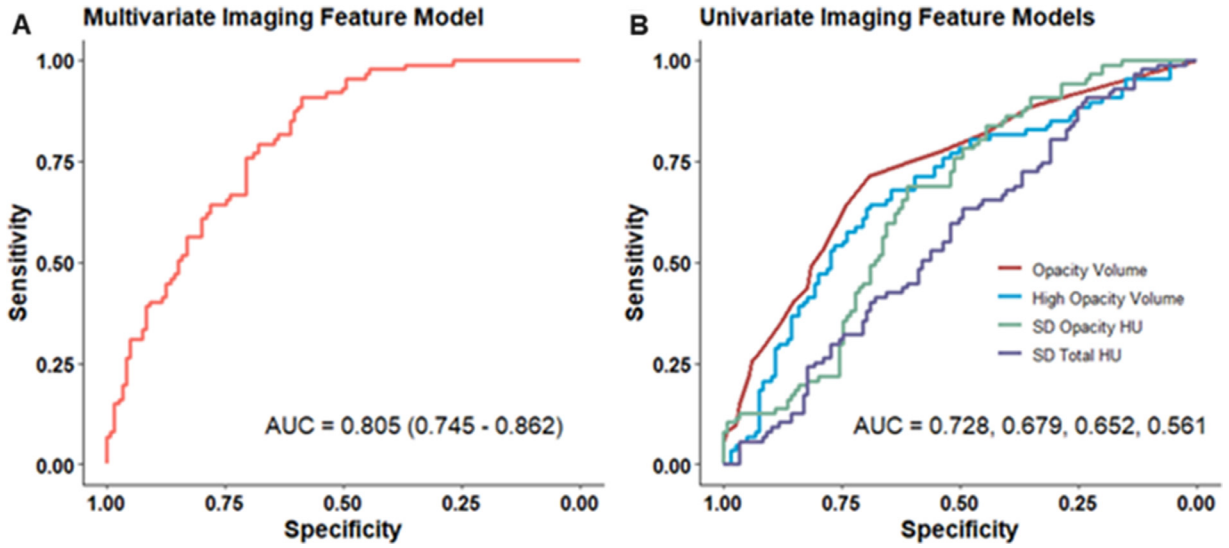



**Figure 3.** AI-segmented imaging features for use in prediction of COVID-19 status. (**a**) Multivariate imaging model consisting of total opacity volume ($cm^3$), high opacity volume ($cm^3$), standard deviation of opacity Hounsfield units, and total standard deviation of all Hounsfield units. All variables were significant in multiple logistic regression model ($p < 0.05$). (**b**) Individual features used in the multivariate model and their individual diagnostic performance for COVID-19 diagnosis. Opacity volume, followed by high opacity volume, had the largest predictive power for COVID-19 diagnosis. AI, Artificial intelligence; AUC, area under curve; COVID-19, Coronavirus disease 2019; HU, Hounsfield units; SD, standard deviation. (Color version of figure is available online.)

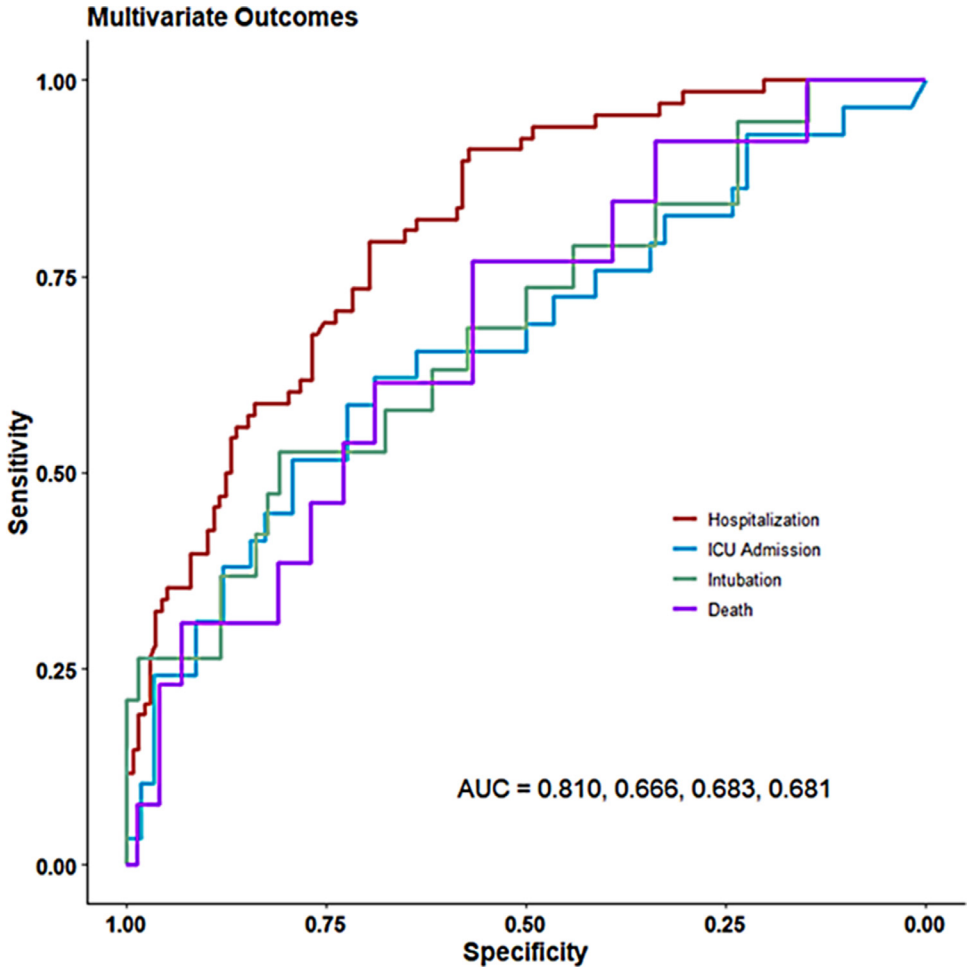



**Figure 4.** Multivariate logistic regression modelling of outcomes used in this study. A combination of variables (Opacity Volume, High Opacity Volume, SD Opacity HU, SD Total HU) derived from significant predictors of COVID-19 status predict hospitalization, ICU admission, intubation and death.

AUC, area under curve; COVID-19, Coronavirus disease 2019; HU, Hounsfield units; ICU, intensive care unit. (Color version of figure is available online.)

**TABLE 2. Diagnostic Parameters of the Most Accurate Thresholds for Inpatient Outcomes. AI Airspace Score Thresholds Have a High Specificity and NPV for Identifying Patients at Risk of Morbidity and Mortality**

| AI Airspace Opacity Score ≥ 8 and Hospitalization (N = 241) | | | |
|---|---|---|---|
| | Hospitalization | No Hospitalization | % Hospitalized |
| Opacity Score ≥8 | 45 | 28 | 61.6% |
| Opacity Score <8 | 26 | 143 | 15.4% |
| Accuracy | 0.777 (0.724-0.829) | Odds Ratio | 8.8 (4.7-16.5) |
| Sensitivity | 0.634 (0.511-0.745) | PPV | 0.616 (0.505-0.728) |
| Specificity | 0.836 (0.772-0.888) | NPV | 0.846 (0.792-0.901) |
| **AI Airspace Opacity Score ≥ 9 and ICU Admission (N = 241)** | | | |
| | ICU Admission | No ICU Admission | % ICU |
| Opacity Score ≥9 | 18 | 49 | 26.9% |
| Opacity Score <9 | 13 | 162 | 7.4% |
| Accuracy | 0.744 (0.689-0.799) | Odds Ratio | 4.58 (2.1-9.8) |
| Sensitivity | 0.581 (0.391-0.755) | PPV | 0.269 (0.163-0.375) |
| Specificity | 0.768 (0.705-0.823) | NPV | 0.926 (0.887-0.965) |
| **AI Airspace Opacity Score ≥ 12 and Intubation (N = 241)** | | | |
| | Intubation | No Intubation | % Intubation |
| Opacity Score ≥12 | 9 | 28 | 24.3% |
| Opacity Score <12 | 11 | 194 | 5.4% |
| Accuracy | 0.839 (0.793-0.885) | Odds Ratio | 5.67 (2.07-9.95) |
| Sensitivity | 0.450 (0.231-0.685) | PPV | 0.243 (0.105-0.381) |
| Specificity | 0.874 (0.823-0.915) | NPV | 0.946 (0.915-0.977) |
| **AI Airspace Opacity Score ≥ 13 and Mortality (N = 241)** | | | |
| | Dead | Alive | % Mortality |
| Opacity Score ≥13 | 4 | 29 | 12.1% |
| Opacity Score <13 | 10 | 199 | 4.8% |
| Accuracy | 0.839 (0.793-0.885) | Odds Ratio | 2.75 (0.852-0.844) |
| Sensitivity | 0.286 (0.084-0.581) | PPV | 0.121 (0.010-0.233) |
| Specificity | 0.873 (0.822-0.913) | NPV | 0.952 (0.923-0.981) |

AI, Artificial intelligence; ICU, Intensive care unit; NPV, Negative predictive value; PPV, Positive predictive value.

Comparison of predictive ability with the literature at large is a challenging task due to the heterogeneity of methods, preponderance of public dataset usage and transfer learning, and a risk of bias (1). Fewer studies still have investigated an interpretable AI severity score from chest CT for both diagnosis and prognosis, but among those with similar aims the correlation coefficients are usually high between the experts and AI (0.87-0.97) (7, 24, 25). Prognostication often falls somewhat less accurate with AUCs between 0.75 and 0.90 reported (26, 27). Univariate severity score AUCs in this range should be expected as other clinical variables (age, immunosuppression, etc.) contribute to disease progression and mortality in patients with COVID-19. A recent study found AUCs of 0.70-0.77 for inpatient outcomes by use of deep learning, which corroborates with our results (28). It may be possible to achieve higher AUCs using radiologists supervised transfer learning (29). It is likely that the univariate prediction strength of current AI methods lies within this range, but we suggest that our study stands out in this cohort due to the use of interpretable AI derived classification schemes.

Empirically derived opacity score thresholds improve on the accuracy and predictive ability of COVID-19-related inpatient outcomes (12, 30, 31). Many clinicians and patients are concerned about the next large decision points in COVID-19 clinical care, and airspace opacity scoring accurately prognosticates patient risk with negative predictive values > 90%; below 8 for hospitalization, below 9 for ICU admission, below 12 for intubation, and below 13 for death. For a patient with an airspace severity score of 2, 5, or 10, a physician could relate that the probability of death to be low at <10%. Conversely, for a hospitalized patient with an airspace opacity score of 17 and approaching escalation of care, a physician could quote upwards of 25% risk of intubation and 20% all-comers mortality when discussing goals of care. Furthermore, AI airspace opacity scoring can inform clinicians, patients, and hospital systems of length of stay and duration of high-level of care including invasive ventilation duration and ICU bed occupancy. Bracketing airspace opacity into quintiles demonstrates a clear upward trend in hospitalization duration, ICU duration, and intubation duration as a function of severity. Physicians could once again counsel a patient with a score of 15 and approaching intubation to expect an invasive ventilation duration of 10 days on average, albeit with a large degree of variation.

CT scans are obtained at other points in admission besides the initial encounter stages and with other possible viral pathologies. While the strategy employed in this study utilizes

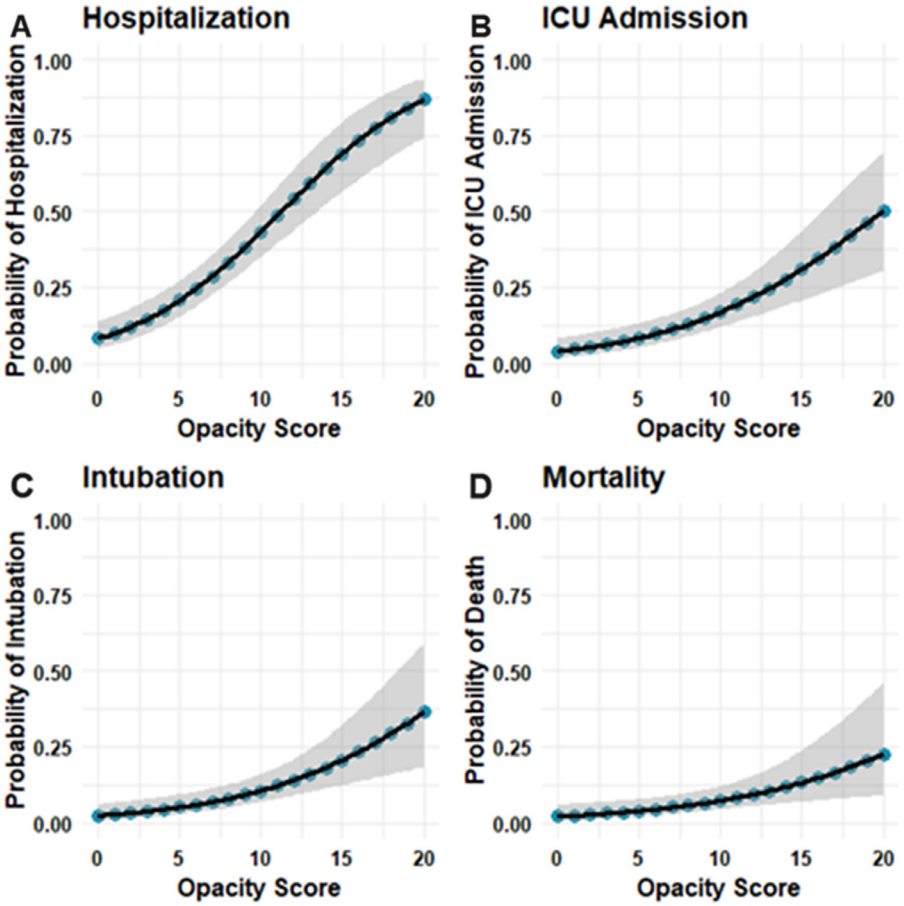



**Figure 5.** Probability of inpatient outcomes among as a logistic function of AI opacity score. The probability of an inpatient event follows an exponential function. AI, Artificial intelligence. (Color version of figure is available online.)

a cross-sectional time point (emergency department admission predicated around SARS-CoV-2 PCR testing), there is a lack of information on if follow-up scores would predict morbidity and mortality as anticipated. The authors find likely that a change in clinical situation should result in a differential rate of outcomes, but there is dearth of follow-up CT scans during hospital admission. At the present time we are unable to conclude if and how the severity score predictions would change over the course of the admission, and instead recommend interpretation of prognostics in the setting of early workup of disease. Regarding other causes of atypical pneumonia, CT has been well described in the evaluations of other viral pneumonias (32, 33). Various deep learning algorithms have attempted to differentiate between COVID-19 and other viral pneumonias; however, the authors argue that with widespread SARS-CoV-2 testing availability this is less of a concern (25). Future study should investigate patients who had subsequent cross-sectional imaging during the hospital course and assess for changes in prognostic value.

## LIMITATIONS

Limitations of this study include the single institution, retrospective nature of this study spanning multiple iterations of COVID-19 waves, vaccines, strains, and best practices. There is no current data to suggest how clinicians might approach this potential confounding aspect (i.e., radiologic findings of the Delta vs Omicron variant, vaccinated vs non-vaccinated vs booster received, etc). Additionally, this study was not powered to evaluate concurrent demographic and comorbidities as risk factors or effect modifiers as those were considered secondary endpoints. Further study is needed to develop more accurate risk modelling in the context of previously identified demographic and clinical variables. The patients enrolled in this study are also subject to selection bias by the criteria of having received a chest CT upon presentation. Patients who receive a CT scan in the ED more likely represent a population with more severe presenting illness, which may inflate the average airspace opacity score among COVID-19 positive patients. Severe outcomes in the COVID-19 group were sparse. A larger multi-institutional cohort is needed with more outcomes, for which this study will serve as the basis for a second power analysis.

Importantly, interpreting airspace opacities in the context of COVID-19 patients is murky, as the type of airspace opacity is not discriminated against by the AI program. For instance – ground-glass opacities, "tree-in-bud" pattern, and patchy consolidation is found in many patients without

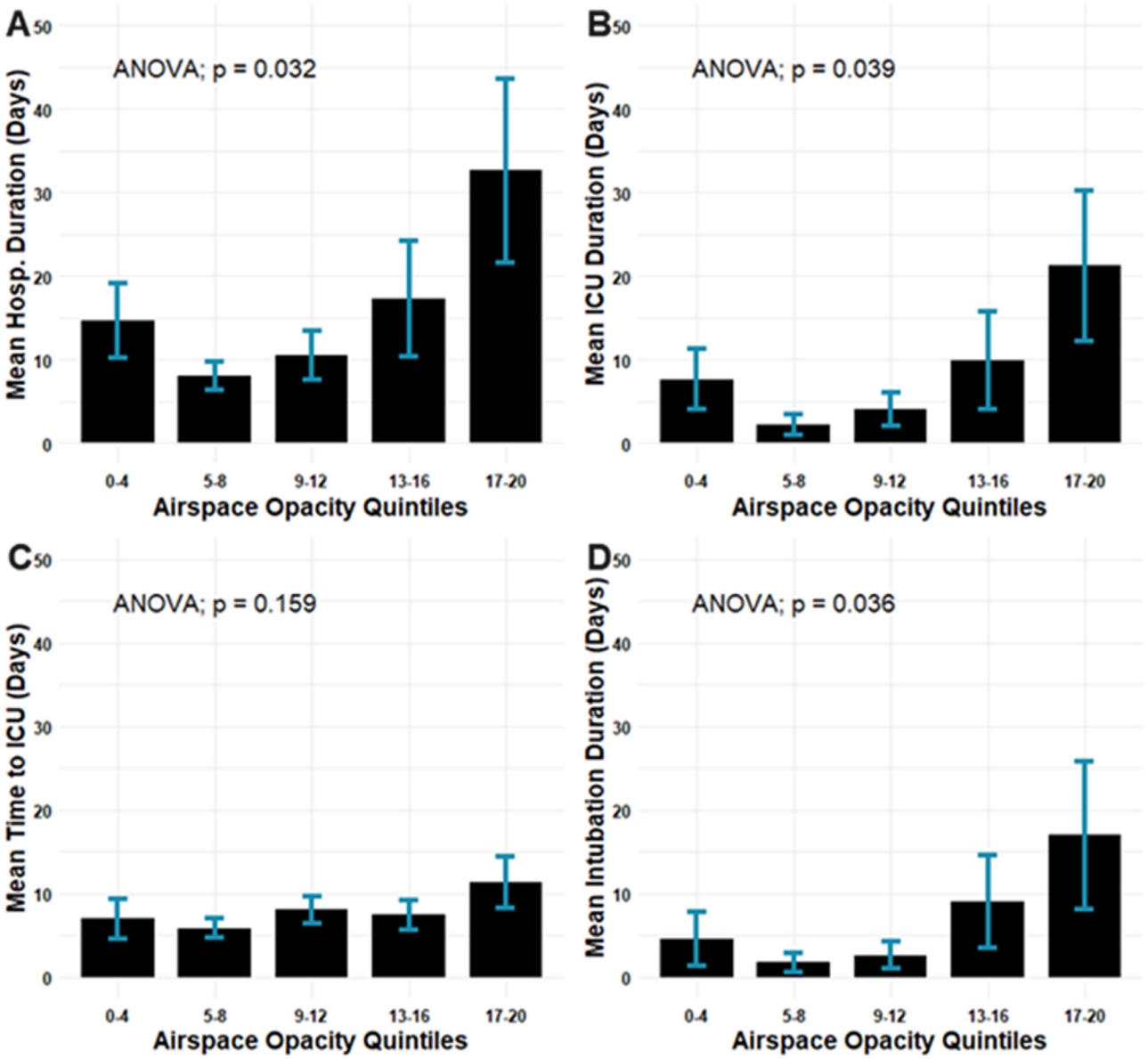



**Figure 6.** Time to event and inpatient duration analysis among hospitalized COVID-19 patients using airspace opacity score quintiles with reported means and standard errors. Mean hospitalization duration (**a**), ICU duration (**b**), and intubation (**d**) duration were associated with increased AI airspace opacity scores. Time from hospital admission to ICU admission (**c**) was not significantly associated with AI airspace opacity scores. AI, Artificial intelligence; ANOVA, analysis of variance; COVID-19, Coronavirus disease 2019; Hosp, hospitalization; ICU, intensive care unit. (Color version of figure is available online.)

pulmonary disease or in non-COVID-19 viral pneumonia, but would count as a positive airspace opacity in patients with or without COVID-19 in this study (2, 3, 34). Certain systems have been invented to classify COVID vs non-COVID pneumonia, but at the time of this article this is still an evolving science. A best practice would include performing a COVID-19 test before cross-sectional imaging to clarify pre-test probability (32).

## CONCLUSIONS

The use of AI segmented quantitative airspace severity scoring is an accurate diagnostic and prognostic tool for COVID-19. The AI algorithm adequately quantifies burden of disease in COVID-19 patients and can provide a service which would otherwise be too time consuming for radiologists and clinicians. The AI scoring output is also easily interpretable, explaining the outputs of a convolutional neural network with relatively little previous knowledge required. Extra value is also provided to clinicians on the risk of progression of disease to their patients, which may change management and influence goals of care discussion. Further study will focus on multivariate predictive outcomes analysis with less emphasis on interobserver agreement.

## FUNDING

Funding for this study was provided by Siemens Healthineers.

## REFERENCES


1. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021; 3:199–217.
2. Goyal N, Chung M, Bernheim A, et al. Computed tomography features of Coronavirus Disease 2019 (COVID-19): a review for radiologists. J Thorac Imaging 2020; 35:211–218.
3. Simpson S, Kay FU, Abbara S, et al. Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA - Secondary Publication. J Thorac Imaging 2020; 35:219–227.
4. Smereka P, Anthopolos R, Latson Jr. LA, et al. Using lung base Covid-19 findings to predict future disease trends and new variant outbreaks: study of first New York City (NYC) outbreak. Acad Radiol 2021; 28:1645–1653.
5. Hurt B, Kligerman S, Hsiao A. Deep learning localization of pneumonia: 2019 Coronavirus (COVID-19) outbreak. J Thorac Imaging 2020; 35: W87–W89.
6. Lessmann N, Sanchez CI, Beenen L, et al. Automated assessment of COVID-19 reporting and data system and Chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. Radiology 2021; 298:E18–E28.
7. Goncharov M, Pisov M, Shevtsov A, et al. CT-Based COVID-19 triage: deep multitask learning improves joint identification and severity quantification. Med Image Anal 2021; 71:102054.
8. Lassau N, Ammari S, Chouzenoux E, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat Commun 2021; 12(1):634.
9. Cai W, Liu T, Xue X, et al. CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. Acad Radiol 2020; 27:1665–1678.
10. Mader C, Bernatz S, Michalik S, et al. Quantification of COVID-19 opacities on Chest CT - evaluation of a fully automatic AI-approach to noninvasively differentiate critical versus noncritical patients. Acad Radiol 2021; 28:1048–1057.
11. Bernheim A, Mei X, Huang M, et al. Chest CT findings in Coronavirus Disease-19 (COVID-19): relationship to duration of infection. Radiology 2020; 295(3):200463.
12. Chaganti S, Grenier P, Balachandran A, et al. Automated quantification of CT patterns associated with COVID-19 from Chest CT. Radiol Artif Intell 2020; 2:e200048.
13. Bouchareb Y, Moradi Khaniabadi P, Al Kindi F, et al. Artificial intelligence-driven assessment of radiological images for COVID-19. Comput Biol Med 2021; 136:104665.
14. Tang Z, Zhao W, Xie X, et al. Severity assessment of COVID-19 using CT image features and laboratory indices. Phys Med Biol 2021; 66 (3):035015.
15. Ferreira Junior JR, Cardenas DAC. The potential role of radiogenomics in precision medicine for COVID-19. J Thorac Imaging 2021; 36(3):W34.
16. Timaran-Montenegro DE, Torres-Ramirez CA, Morales-Jaramillo LM, et al. Computed tomography-based lung residual volume and mortality of patients with Coronavirus Disease-19 (COVID-19). J Thorac Imaging 2021; 36:65–72.
17. Fu L, Li Y, Cheng A, et al. A novel machine learning-derived radiomic signature of the whole lung differentiates stable from progressive COVID-19 infection: a retrospective cohort study. J Thorac Imaging 2020; 35:361–368.
18. Kim FD-VaB. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*2017;
19. Wasilewski PG, Mruk B, Mazur S, et al. COVID-19 severity scoring systems in radiological imaging - a review. Pol J Radiol 2020; 85:e361–e368.
20. Kwon YJF, Toussie D, Finkelstein M, et al. Combining initial radiographs and clinical variables improves deep learning prognostication in patients with COVID-19 from the emergency department. Radiol Artif Intell 2021; 3:e200098.
21. Li K, Fang Y, Li W, et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). Eur Radiol 2020; 30:4407–4416.
22. Yang R, Li X, Liu H, et al. Chest CT severity score: an imaging tool for assessing severe COVID-19. Radiol Cardiothorac Imaging 2020; 2: e200047.
23. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med 2020; 8(7):501.
24. Wang S, Zha Y, Li W, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020; 56 (2):2000775.
25. Wang G, Liu X, Shen J, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. Nat Biomed Eng 2021; 5:509–521.
26. Homayounieh F, Bezerra Cavalcanti Rockenbach MA, Ebrahimian S, et al. Multicenter assessment of CT pneumonia analysis prototype for predicting disease severity and patient outcome. J Digit Imaging 2021; 34:320–329.
27. Feng Z, Shen H, Gao K, et al. Machine learning based on clinical characteristics and chest CT quantitative measurements for prediction of adverse clinical outcomes in hospitalized patients with COVID-19. Eur Radiol 2021; 31:7925–7935.
28. Lee EH, Zheng J, Colak E, et al. Deep COVID DeteCT: an international experience on COVID-19 lung detection and prognosis using chest CT. NPJ Digit Med 2021; 4(1):11.
29. Hurt B, Rubel MA, Masutani EM, et al. Radiologist-supervised transfer learning: improving radiographic localization of pneumonia and prognostication of patients with COVID-19. J Thorac Imaging 2021; 37(2):90–99.
30. Huang L, Han R, Ai T, et al. Serial quantitative chest CT assessment of COVID-19: a deep learning approach. Radiol Cardiothorac Imaging 2020; 2:e200075.
31. Shen C, Yu N, Cai S, et al. Quantitative computed tomography analysis for stratifying the severity of Coronavirus Disease 2019. J Pharm Anal 2020; 10:123–129.
32. Borges da Silva Teles G, Kaiser Ururahy Nunes Fonseca E, Yokoo P, et al. Performance of chest computed tomography in differentiating Coronavirus Disease 2019 from other viral infections using a standardized classification. J Thorac Imaging 2021; 36:31–36.
33. Zarei F, Jalli R, Iranpour P, et al. Differentiation of Chest CT findings between influenza pneumonia and COVID-19: interobserver agreement between radiologists. Acad Radiol 2021; 28:1331–1338.
34. Pezzutti DL, Wadhwa V, Makary MS. COVID-19 imaging: diagnostic approaches, challenges, and evolving advances. World J Radiol 2021; 13:171–191.


## SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.acra.2022.03.023.