Value of quantitative airspace disease measured on chest CT and chest radiography at initial diagnosis compared to clinical variables for prediction of severe COVID-19

Hae-Min Jung,[a] Rochelle Yang,[a] Warren B. Gefter,[a] Florin C. Ghesu,[b] Boris Mailhe,[b] Awais Mansoor,[b] Sasa Grbic,[b] Dorin Comaniciu,[b] Sebastian Vogt,[c] and Eduardo J. Mortani Barbosa Jr.[a,*]

[a]University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, United States

[b]Siemens Healthineers, Digital Technology and Innovation, Princeton, New Jersey, United States

[c]Siemens Healthineers, X-Ray Products, Malvern, Pennsylvania, United States

Abstract

Purpose: Rapid prognostication of COVID-19 patients is important for efficient resource allocation. We evaluated the relative prognostic value of baseline clinical variables (CVs), quantitative human-read chest CT (qCT), and AI-read chest radiograph (qCXR) airspace disease (AD) in predicting severe COVID-19.

Approach: We retrospectively selected 131 COVID-19 patients (SARS-CoV-2 positive, March to October, 2020) at a tertiary hospital in the United States, who underwent chest CT and CXR within 48 hr of initial presentation. CVs included patient demographics and laboratory values; imaging variables included qCT volumetric percentage AD (POv) and qCXR area-based percentage AD (POa), assessed by a deep convolutional neural network. Our prognostic outcome was need for ICU admission. We compared the performance of three logistic regression models: using CVs known to be associated with prognosis (model I), using a dimension-reduced set of best predictor variables (model II), and using only age and AD (model III).

Results: 60/131 patients required ICU admission, whereas 71/131 did not. Model I performed the poorest (AUC = 0.67 [0.58 to 0.76]; accuracy = 77%). Model II performed the best (AUC = 0.78 [0.71 to 0.86]; accuracy = 81%). Model III was equivalent (AUC = 0.75 [0.67 to 0.84]; accuracy = 80%). Both models II and III outperformed model I (AUC difference = 0.11 [0.02 to 0.19], $p = 0.01$; AUC difference = 0.08 [0.01 to 0.15], $p = 0.04$, respectively). Model II and III results did not change significantly when POv was replaced by POa.

Conclusions: Severe COVID-19 can be predicted using only age and quantitative AD imaging metrics at initial diagnosis, which outperform the set of CVs. Moreover, AI-read qCXR can replace qCT metrics without loss of prognostic performance, promising more resource-efficient prognostication.

 [DOI: 10.1117/1.JMI.9.3.034003]

Keywords: COVID-19; artificial intelligence; chest imaging; prognosis.

Paper 22011GR received Jan. 27, 2022; accepted for publication May 31, 2022; published online Jun. 17, 2022.

1 Introduction

Since its emergence in December 2019 in Wuhan, China, the global outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated disease—COVID-19—has accrued over 2160 million confirmed cases and over 4,500,000 confirmed deaths.[1] To reduce

*Address all correspondence to Eduardo J. Mortani Barbosa Jr., eduardo.barbosa@pennmedicine.upenn.edu

the overall burden of COVID-19 on healthcare systems, especially in the face of additional variants, diagnosis and prognostication should be accurate, efficient, and accessible. For diagnosis, the reference standard is reverse-transcriptase polymerase chain reaction (RT-PCR). This can be aided by imaging modalities such as computed tomography (CT) and chest radiography (CXR), which have been studied for diagnosis and for monitoring progression using AI methods.[2–6] The most common thoracic imaging manifestations of COVID-19 pneumonia are reported as peripheral, bilateral ground glass opacities with or without consolidation or visible interlobular lines.[7,8]

Because COVID-19 can vary widely in initial presentation and disease progression—from asymptomatic to multiorgan failure and death—much research has focused on finding clinical and imaging prognostic indicators.[9,10] These studies have used a variety of endpoints, such as mortality, intensive care (ICU) admission, length-of-stay, and disease severity rating.[10,11] Some studies have identified clinical and demographic factors associated with more severe outcomes, and prognostic algorithms based on these variables have been proposed for clinical use.[12] For prognosis prediction based on imaging, visual semiquantitative scoring systems of CT and CXR have been proposed.[3,13,14] Others have applied AI to CT images for prognosis prediction.[15–18] Due to the large number of potential predictors, machine learning/artificial intelligence methods are used extensively in these studies for feature extraction and classification, with potential for more rapid results over human readers.[11]

However, there is scarce literature on direct comparisons of relative prognostic abilities between clinical, quantitative CXR, and quantitative CT variables using multiparametric statistical models. Due to inconsistencies in clinical and laboratory data gathered and availability of various imaging modalities across different healthcare systems, such a comparison is crucial in working toward not just an accurate prognostic model but a maximally resource-efficient one.

To compare these many variables efficiently, feature selection is required. Embedded methods, such as random forest, are popular in feature selection due to their robustness to noise and resistance to overfitting.[19] However, both wrapper and embedded methods become computationally intensive with higher feature datasets, lack interpretability as "black box" methods, and have been found biased by correlated features.[20] To address these weaknesses, embedded methods can be combined with filter methods, such as minimum redundance-maximum relevance (mRMR) which quickly culls variables by their relevance while penalizing redundancy to other variables.[21] mRMR has the additional advantage of interpretability of the features filtered. Developed for and applied to genomics analysis, mRMR in combination with advanced embedded/wrapper methods has been found more accurate than using classifiers alone, with lower computational cost.[22] Here, we use mRMR in conjunction with random forest for simultaneous feature selection and inferences on the relative prognostic strength of variables.

Our goals in this study were to assess whether severe COVID-19, which we defined as requiring ICU admission, can be accurately predicted using a set of demographics, clinical, qCXR, and qCT variables at initial presentation. ICU admission was chosen to address serious disease manifestations, including but not limited to respiratory failure. Our secondary goal was to assess if AI-read qCXR was a more cost- and time-efficient, noninferior predictor compared with human-read qCT, acknowledging that a direct measure of any efficiency gains in radiologists' interpretation is beyond the scope of this study. Moreover, we assessed the relative performance of models incorporating subsets of clinical and imaging variables with cross-validation, to rank the most valuable predictors and arrive at a parsimonious optimal model with a minimum set of variables.

2 Materials and Methods

2.1 *Patient and Image Selection*

This single institution, retrospective study obtained IRB approval with waiver of informed consent and was HIPAA compliant. We randomly selected 131 patients with the following inclusion criteria: positive RT-PCR for SARS-CoV-2, a pair of CXR, and chest CT performed within 48 h of each other and within 48 h of presentation, and age older than 18. All scans were obtained

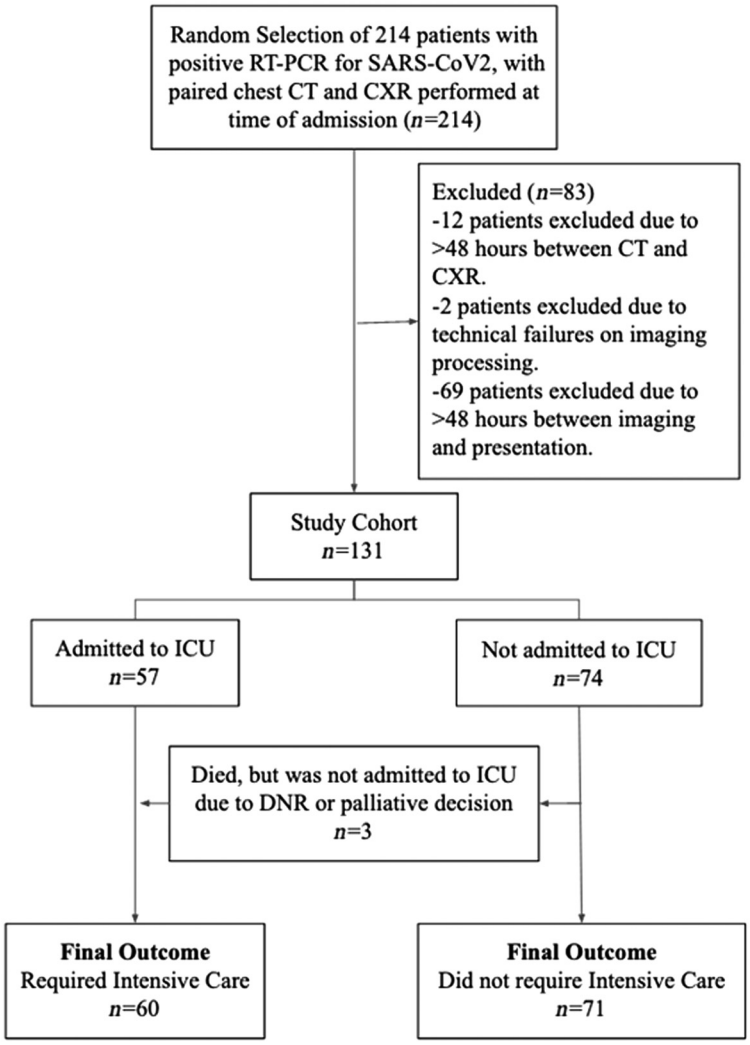

Fig. 1 Flow diagram with inclusion and exclusion criteria in our study cohort ($n = 131$).

within the following date range: March to October, 2020. Figure 1 details study design, inclusion, and exclusion criteria.

CXR and CT images were de-identified using a standard anonymization profile in Sectra PACS and transferred through a secure file exchange to a computational cluster for imaging processing. Table 1 lists all variables included for analysis for each patient.

2.2 *Clinical and Demographic Variables*

We collected the following demographic variables from the EMR (electronic medical record): age, sex, race, employment, and insurance status, smoking/alcohol/drug use history, and marital status. In addition, we collected the following CVs, as close to initial presentation as possible, through EMR chart review:

- Vital signs and biometrics: BMI, heart rate, blood pressure, temperature, and respiratory rate.
- Laboratorial values: white blood cells (WBC), red blood cells (RBC), hemoglobin (Hgb), red cell distribution width (RDW), mean cell hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), platelets (Plt), aspartate aminotransferase (AST), alanine aminotransferase (ALT), creatinine, erythrocyte sedimentation rate (ESR), D-dimer, lactate dehydrogenase (LDH), ferritin, C-reactive protein (CRP), troponin T.

Table 1 Variables by category.

Demographics	Vital signs	Laboratory	AD metrics
Age	Respiratory rate	AST (abnormal)	POa CXR CNN
Cardiovascular comorbidities	Symptom-neurologic	ESR (abnormal)	POa DRR CNN
Gender	Temperature	CRP (abnormal)	POv CT ground-truth
Renal comorbidities	BP_abnormal	Creatinine (abnormal)	—
BMI	Symptom-fever	D.Dimer (abnormal)	—
Immunologic comorbidities	Symptom-GI	MCHC (abnormal)	—
Diabetes	Heart rate	ALT (abnormal)	—
Other comorbidities	Symptom-respiratory	RDW (abnormal)	—
Respiratory comorbidities	—	MCV (abnormal)	–
—	—	Hct (abnormal)	—
—	—	Hgb (abnormal)	—
—	—	Ferritin (abnormal)	—
—	—	WBC (abnormal)	—
—	—	MCH (abnormal)	—
—	—	Troponin.T (abnormal)	—
—	—	Plt (abnormal)	—
—	—	RBC (abnormal)	—

- Presenting symptoms: dyspnea, chest pain, fever, loss of taste/smell, other neurologic, and GI symptoms.
- Co-morbidities organized by organ system: immunologic/inflammatory, respiratory, cardiovascular, neurologic, gastrointestinal, genitourinary, and diabetes.

Oxygen saturation as measured by pulse oximetry (SpO2) and partial pressure oxygen on venous or arterial blood gas labs were not used for several reasons. Patients brought in by Emergency Medical Services (EMS) were often already on supplemental oxygen, artificially elevating their SpO2 and pO2. SpO2 especially is heavily dependent on patient activity and position, and two patients with the same degree of poor oxygenation can measure significantly different SpO2 if one is at rest and the other just transferred in or out of bed. Furthermore, venous pO2 and arterial pO2 carry different interpretations (tissue oxygen use and cardiovascular oxygen output, respectively), and many patients lacked one or both. Like SpO2, pO2 is also confounded by any use of supplemental oxygen.

Laboratorial values were binarized to 0/1 (1 = abnormal, 0 = normal) with missing values imputed to 0 = normal due to a lack of established evidence of the relative impact of different value ranges. No demographics, symptoms, or comorbidities datum was missing for any of the 131 patients, and these were also binarized to 0/1 (1 = comorbidities/symptom present, 0 = comorbidities/symptom absent).

2.3 *Airspace Disease Quantification on CXR and Chest CT*

All CXRs were performed with AP (anterior posterior) technique and digital acquisition. Chest CTs were performed with variable protocols (on Siemens or GE scanners); however, every scan included contiguous 1-mm axial slices without interslice gap, obtained with iterative reconstruction using high spatial resolution algorithms (tailored for lung evaluation).

Airspace disease (AD) was quantified on CT and on CXR utilizing a combination of expert human annotations on CT and an AI method on CXR, detailed on a previous publication.[23]

Two deep convolutional neural networks (CNNs) are used, one for lung segmentation and one for segmentation of airspace opacity from chest DRR and x-ray. Both networks are based on a residual U-net architecture, with the encoder layers initialized from the ResNet18 architecture.[24,25] This results in a nine-layer deep network with 64, 64, 128, 256, and 512 feature maps in the encoding layers (following the ResNet18 structure). Skip connections based on feature map concatenation are used. The final layer is a mapping to two-channel output, describing foreground (i.e., airspace opacity area) and background.

CT images have been annotated at raw resolution and without preprocessing. DRRs are generated as integrals over synthetic projection lines through the CT volume under a parallel projection geometry. Using a deep CNN, we enhance the resolution of the resulting DRR to achieve isotropic resolution that is higher, closer to typical chest radiographs (CXR). DRRs and x-rays (during inference) are feed to both segmentation networks at a resolution of 512×512 pixels. More details can be found in Ref. 23. In terms of preprocessing, for x-ray processing during inference we use the robust intensity normalization technique described in Ref. 26.

Figure 2 shows our method. Figure 3 provides an example of AD quantification on CT. Figure 4 demonstrates a patient example of multiple types of AD quantification according to our method, on CT, CXR, and DRR, by the CNN algorithm, validated by human expert readers.

The metrics used to assess the severity of AD are as follows:

- Percentage of opacity–volume (POv): The POv is measured on CT scans and quantifies the percent volume of the lung parenchyma that is affected by AD

$$\mathrm{POv} = 100 \times \frac{\text{Volume of Airspace Disease}}{\text{Total Lung Parenchyma Volume}}.$$

- Percentage of opacity – area (POa): The POa is measured on CXRs and quantifies the percent area of the lung parenchyma that is affected by AD:

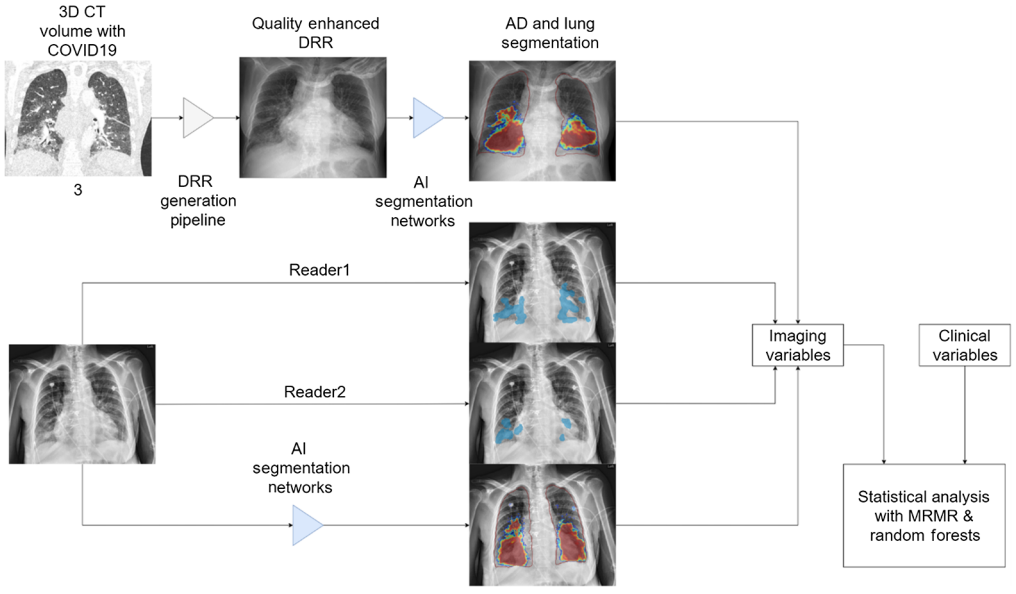

Fig. 2 Schematic illustration of our method, which generates quantitative AD annotation on CXR by a CNN algorithm, leveraging paired CT POv AD (ground-truth) and validated by two independent expert readers. The set of imaging variables were fed into our statistical models, in conjunction with CVs. Light blue color reflects expert human annotation of AD (two expert readers). Color coded heat map (predominantly red) reflects the CNN prediction of AD, with red in the center representing the highest confidence level, and blue at the edges the lowest.

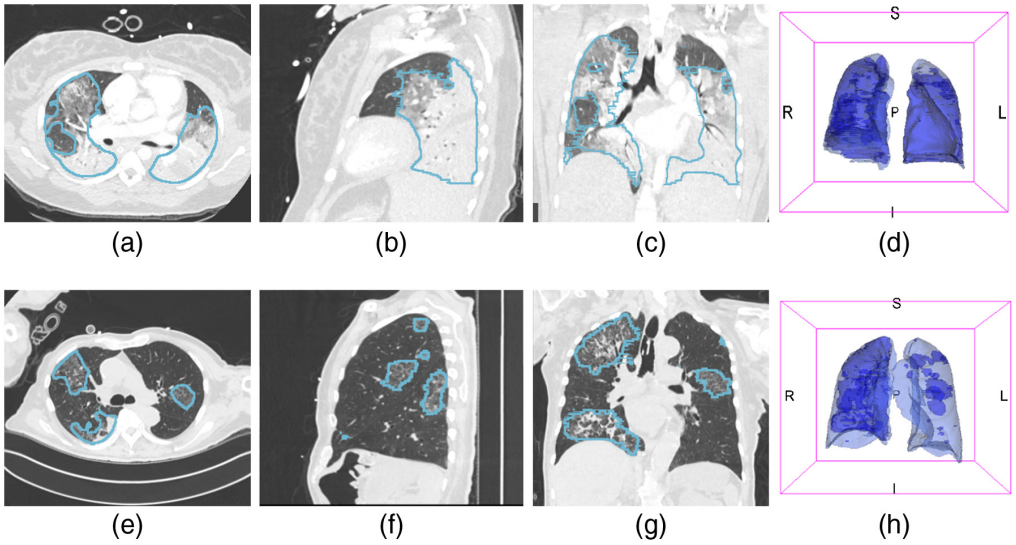

Fig. 3 CT derived three-dimensional volume (POv) airspace quantification (ground truth), on (a, e) axial, (b, f) sagittal, (c, g) coronal MPRs, and on (d, h) VR masks, in two patients in our cohort. Light blue lines reflect the contours of the segmented AD.

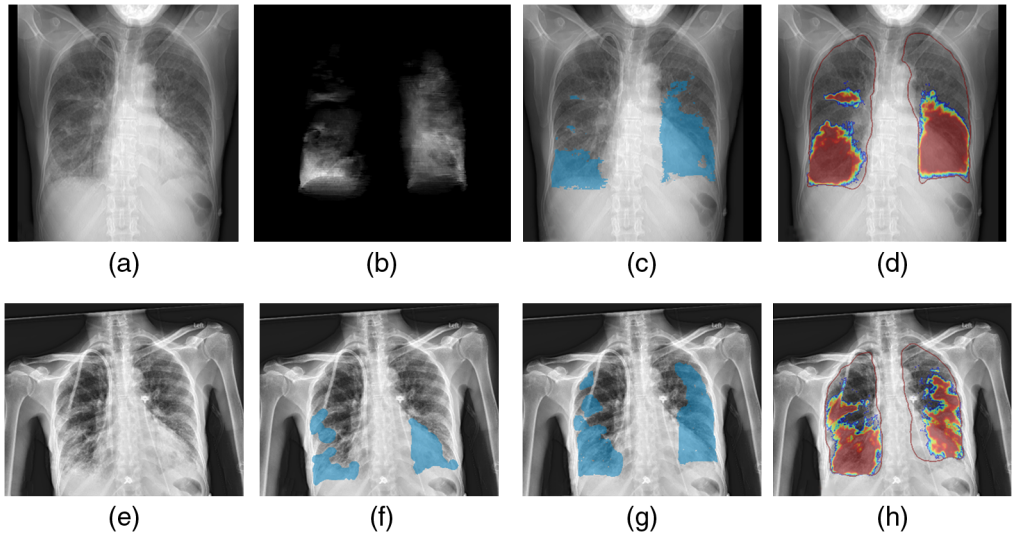

Fig. 4 (a) DRR, (b) AP intensity thickness AD mask from CT, (c) DRR with projected AD mask, (d) DRR CNN prediction, (e) CXR, (f), (g) CXR expert readers for validation, (h) CXR CNN prediction, in a patient example in our cohort. Light blue color reflects expert human annotation of AD [(f) and (g)] two expert readers]. Light blue color in (c) reflects the DRR with projected AD mask from CT ground-truth. Color coded heat map (predominantly red) reflects the CNN prediction of AD, with red in the center representing the highest confidence level, and blue at the edges the lowest (d and h).

$$\mathrm{POa} = 100 \times \frac{\text{Area of Airspace Disease}}{\text{Total of Lung Parenchyma Area}}.$$

The Appendix describes characteristics of AI CNN training datasets.

2.4 *Outcome*

A binary outcome of requiring ICU was chosen, such that all deaths were counted under requiring ICU, and nonadmissions or admissions to a regular ward were counted as not requiring ICU.

All deaths on chart review that occurred out of the ICU were due to do-not-resuscitate orders or patient family decisions for palliative care. Otherwise, these patients would have required ICU due to their clinical condition. This outcome also allows our models to capture the risk of all serious outcomes, not just ventilation or intubation. Of the 131 patients, 71 required admission to the ICU, and 60 did not.

2.5 *Statistical Analysis*

Dimension reduction was conducted by first filtering all variables using a classic mRMR ensemble filter in the *mRMRe* package in R, with the main outcome described already as the target variable.[27] 100 random forests, each with 1000 trees, were then fit on the top 10 variables from mRMR filter using the *randomForest* package in R.[28] Mean accuracy decrease (MAD) importance rankings were averaged across the runs, and the top few variables were statistically tested and fit using logistic regression. Logistic regression was chosen as the classification algorithm due to better interpretability of individual predictors over random forest modeling. The area under the curve (AUC) of a receiver operating curve (ROC) for the model was used to assess prognostic value of the logistic model using leave-one-out cross-validation (LOOCV) approach using the *cvAUC* package in conjunction with the *pROC* package in R.[29,30] In the LOOCV approach, the data were partitioned into folds where one participant's data had been withheld—thus, the number of folds was equal to the number of participants. For each fold, the empiric AUC was calculated, and an LOOCV AUC was estimated by the mean of these empiric AUC. The ROC was plotted using a vertical average from the LOOCV runs. Prediction accuracy for each logistic model was also calculated using LOOCV with the *cvAUC* package.

Three logistic models were fit and compared: (1) A base model based solely on CVs found consistently associated with COVID-19 prognosis in the literature (model I); (2) a model fit using the highest-yield variables from dimension reduction (model II); and (3) a model fit only on age and imaging (qCT or qCXR) variables (model III). Parallel versions of models II and III were also fit, replacing POv CT with POa from AD on CXR calculated by CNN. The AUC of models I, II, and III was compared pairwise using the paired Delong Test for correlated ROC, included in the *pROC* package. The paired Delong Test was also used to compare the AUC of the parallel models of models II and III. ANOVA was not conducted as the ROC for these models are correlated, and repeated measures ANOVA was not possible due to the inability to calculate subject-specific AUC.

These statistical analyses were conducted on a single computer with 16 GB of RAM and an i7-4790 CPU. We define statistical significance at $p < 0.05$.

3 Results

Of the 131 total patients, 48% (63) were male, and 52% (68) female, with a mean age of 60.3 and standard deviation of 18.0. Of the 60 patients who required ICU care, 48% (29) were male, 52% (31) were female, with a mean age of 67.1 and standard deviation of 16.4. Of the 71 patients who did not require ICU care, 48% (34) were male, 52% (37) were female, with a mean age of 54.4 and standard deviation of 17.3 (Table 2).

In the dimension reduction, age, respiratory rate, abnormal creatinine, POv on CT, and abnormal MCHC were identified as the top five variables by random forest among the top 10 variables filtered by mRMR (Table 3).

The CVs only model (model I) had the poorest performance and accuracy: diabetes, cardiovascular comorbidities, respiratory comorbidities, and BMI were insignificant, only age was significant (AUC = 0.67 [0.58 to 0.76]; accuracy = 77%). In the logistic model fit using variables selected from mRMR and random forest dimension reduction (model II), all variables but respiratory rate were significant predictors of ICU (AUC = 0.78 [0.70 to 0.86]; accuracy = 81%). A parsimonious model with only age and affected volume on CT (model III) performed similarly without need of laboratory tests (AUC = 0.75 [0.67 to 0.84]; accuracy = 80%; AUC difference to model II = 0.03 (95%: −0.02 to 0.07, $p = 0.27$)). Both models II and III

Table 2 Demographics of cohort ($n = 131$).

	Total ($n = 131$)	ICU ($n = 60$)	Non-ICU ($n = 71$)
Age (mean ± SD, range)	**60.3 ± 18.0 (20 to 98)**	**67.2 ± 16.4 (25 to 97)**	**54.4 ± 17.3 (20 to 98)**
Gender			
Male	48% (63/131)	48% (29/60)	48% (34/71)
Female	52% (68/131)	52% (31/60)	52% (37/71)
Race/ethnicity			
White non-Hispanic	20% (26/131)	20% (12/60)	17% (14/71)
White Hispanic	0.8% (1/131)	1% (1/60)	0% (0/71)
Black or African American	69% (91/131)	68% (41/60)	70% (50/71)
Black Hispanic	0.8% (1/131)	0% (0/60)	1% (1/71)
Asian	1.5% (2/131)	3% (2/60)	0% (0/71)
East Indian	0.8% (1/131)	0% (0/60)	1% (1/71)
Other/unknown	6% (8/131)	7% (3/60)	7% (5/71)
Smoking status			
Never smoked	59% (77/131)	45% (27/60)	70% (50/71)
Former/current smoker	30% (40/131)	40% (22/60)	20% (16/71)
Unknown	11% (14/131)	15% (9/60)	3% (5/71)
Outcome			
Discharged	80% (105/131)	57% (34/60)	100% (71/71)
Ventilation	21% (28/131)	47% (28/60)	0% (0/71)
Deceased	19% (26/131)	43% (26/60)	0% (0/71)

Table 3 Top 10 variables identified by mRMR, ranked by MAD on random forest (MCHC, mean corpuscular hemoglobin concentration; CRP, C-reactive protein; AST, aspartate aminotransferase).

	mRMR score	MAD score
Age	0.069	20.067
Respiratory rate	0.007	12.823
Abnormal creatinine	0.034	11.216
POv	0.042	10.319
Abnormal MCHC	0.026	7.087
Abnormal CRP	0.025	6.822
Abnormal AST	0.032	5.208
Renal comorbidities	0.012	4.754
Cancer comorbidities	0.018	2.891
Neurologic symptoms	0.010	2.530

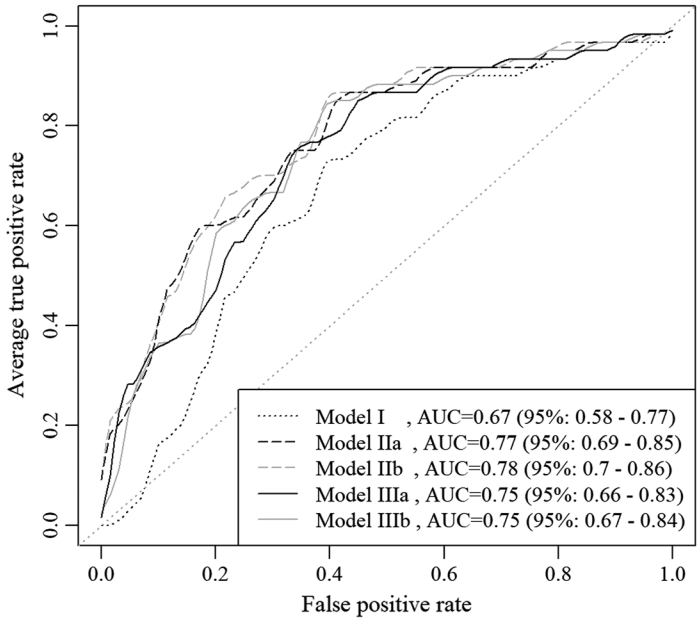

Fig. 5 Vertically averaged LOOCV ROC of models I to III, with parallel models II and III using POa (CXR CNN, model "a") and POv (CT, model "b"), respectively.

had significantly higher AUC compared to model I (AUC difference = 0.11 (95%: 0.02 to 0.19, $p = 0.01$); AUC difference = 0.08 (95%: 0.01 to 0.15, $p = 0.04$), respectively). Neither models II nor III performed significantly differently when POv CT was swapped out for AI CNN-obtained POa on CXR (AUC difference = 0.01 (95%: −0.01 to 0.02, $p = 0.23$); AUC difference = 0.01 (95%: −0.01 to 0.03, $p = 0.45$), respectively) (Figs. 5 and 6). Tables 4 and 5 contain variable coefficient summaries for all models fit as well as results of the Delong tests comparing correlated AUC's.

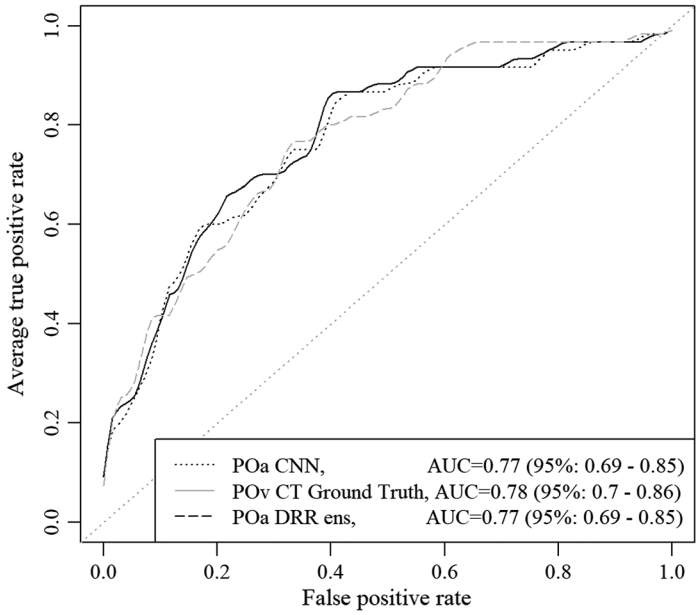

Fig. 6 Vertically averaged LOOCV ROC of model II, with different imaging variables (POa CXR CNN, POv CT, and POa DRR CNN).

Table 4 Individual variable coefficients, confidence intervals, and *p*-values from models I to III (ROC, receiver operating curve; AUC, area under the curve; LOOCV, leave-one-out cross-validation; BMI, body mass index (Kg/m^2); CV comorbidities, cardiovascular comorbidities (0/1 binary, 1 = comorbidities present); MCHC, mean corpuscular hemoglobin concentration (pg.); POv, percent of opacity-volume; CT, computed tomography; POa, percent of opacity-area; DRR CNN, digitally reconstructed radiographs; CNN, convolutional neural network.)

	Model I	Model II a	Model II b	Model III a	Model III b
ROC AUC	0.67 (95%: 0.58 to 0.76)	0.77 (95%: 0.69 to 0.86)	0.78 (95%: 0.70 to 0.86)	0.75 (95%: 0.66 to 0.83)	0.75 (95%: 0.67 to 0.84)
Difference in AUC to model I	—	—	**0.11 (95%:0.02 to 0.19, $p = 0.01$)**	—	**0.08 (95%:0.01 to 0.15, $p = 0.04$)**
Difference in AUC between parallel models	—	0.01 (95%: −0.01 to 0.02, $p = 0.23$)		0.01 (95%: −0.01 to 0.03, $p = 0.45$)	
LOOCV accuracy	77%	81%	81%	80%	80%
Variables					
Age	**0.043 (95%: 0.019 to 0.071, $p = 0.001$)**	**0.041 (95%:0.015 to 0.070, $p = 0.003$)**	**0.041 (95%:0.015 to 0.070, $p = 0.003$)**	**0.052 (95%:0.015 to 0.069, $p < 0.001$)**	**0.051 (95%:0.027 to 0.077, $p < 0.001$)**
BMI	−0.002 (95%: −0.046 to 0.040, $p = 0.91$)	—	—	—	—
Diabetes	−0.006 (95%: −0.87 to 0.85, $p = 0.99$)	—	—	—	—
CV comorbidities	0.51 (95%: −0.32 to 1.34, $p = 0.23$)	—	—	—	—
Respiratory comorbidities	0.53 (95%: −0.26 to 1.35, $p = 0.19$)	—	—	—	—
Respiratory rate	—	0.03 (95%: −0.030 to 0.12, $p = 0.41$)	0.03 (95%: −0.031 to 0.12, $p = 0.42$)	—	—
Abnormal creatinine	—	**0.98 (95%:0.03 to 1.98, $p = 0.046$)**	**1.02 (95%: 0.07 to 2.02, $p = 0.04$)**	—	—
Abnormal MCHC	—	**1.72 (95%:0.25 to 3.48, $p = 0.03$)**	**1.73 (95%:0.28 to 3.47, $p = 0.03$)**	—	—
POv CT	—	—	**3.23 (95%:0.95 to 5.82, $p = 0.009$)**	—	**3.89 (95%:1.70 to 6.41, $p < 0.001$)**
POa CXR CNN	—	**2.90 (95%:0.78 to 5.33, $p = 0.011$)**	—	**3.56 (95%:1.57 to 5.84, $p < 0.001$)**	—

Table 5 Individual variable coefficients, confidence intervals, and *p*-values from model II with different imaging variables (ROC, receiver operating curve; AUC, area under the curve; LOOCV, leave-one-out cross-validation; BMI, body mass index (Kg/m^2), CV comorbidities, cardiovascular comorbidities (0/1 binary, 1 = comorbidities present); CRP, C-reactive protein; POv, percent of opacity-volume; CT, computed tomography; POa, percent of opacity-area; CXR, chest x-ray; DRR, digitally reconstructed radiographs; CNN, convolutional neural network).

	POa CXR CNN	POv CT ground-truth	POa DRR CNN
ROC AUC	0.77 (95%: 0.69 to 0.85)	0.78 (95%: 0.70 to 0.86)	0.77 (95%: 0.69 to 0.85)
Difference in AUC compared to POa CXR CNN	—	0.01 (95%: −0.01 to 0.03, $p = 0.45$)	0.001 (95%: −0.03 to 0.03, $p = 0.94$)
Difference in AUC compared to POv CT ground-truth	—	—	−0.01 (95%: −0.04 to 0.02, $p = 0.63$)
LOOCV accuracy	81%	81%	81%
Age	**0.041 (95%:0.015 to 0.070, $p = 0.003$)**	**0.041 (95%:0.015 to 0.070, $p = 0.003$)**	**0.045 (95%:0.018 to 0.074, $p = 0.002$)**
Respiratory rate	0.03 (95%: −0.030 to 0.12, $p = 0.41$)	0.03 (95%: −0.031 to 0.12, $p = 0.42$)	0.02 (95%: −0.033 to 0.11, $p = 0.53$)
Abnormal creatinine	**0.98 (95%:0.03 to 1.98, $p = 0.046$)**	**1.02 (95%:0.07-2.02, p=0.04)**	**0.92 (95%:0.026 to 1.91, $p = 0.05$)**
Abnormal MCHC	**1.72 (95%:0.25 to 3.48, $p = 0.03$)**	**1.73 (95%:0.28 to 3.47, $p = 0.03$)**	**1.73 (95%:0.30 to 3.47, $p = 0.03$)**
POa CNN	**2.90 (95%:0.78 to 5.33, $p = 0.011$)**	—	—
POv CT GT	—	**3.23 (95%:0.95 to 5.82, $p = 0.009$)**	—
POa DRR CNN	—	—	**3.62 (95%:0.99 to 6.50, $p = 0.009$)**

4 Discussion

Our study demonstrates that COVID-19 requiring ICU can be accurately predicted at presentation with just age and extent of lung parenchyma affected by AD, either % volume on CT (POv) or % area on CXR (POa). The noninferiority of this parsimonious model to the complete model with laboratory values and the relatively poor performance of the prediction model based only on comorbidities suggests that imaging—especially the faster and cheaper AI-read qCXR—may be a more resource-efficient way to assess prognosis than obtaining many laboratorial values. It also suggests that imaging may be a better prognostic predictor than some predictors commonly evaluated in literature, such as BMI, diabetes, or cardiovascular and respiratory comorbidities.

These conclusions are corroborated by both mRMR and random forest dimension reduction, where most nonimaging variables were ranked lower. From mRMR, we can infer that these variables were either not as relevant at predicting ICU admission, contained redundant information, or both. This is reflected in the logistic regression coefficients, where AD coefficients are consistently of higher magnitude than other predictors—this implies that while age is a powerful prognosticator in the elderly, in younger COVID-19 patients their

AD on imaging is the strongest prognosticator of ICU admission. Moreover, the coefficients suggest that older patients tend to require ICU admission with less AD than younger patients.

Furthermore, it also shows that AI-derived AD quantification on CXR is as strong a prognostic predictor of severe COVID-19 as AD quantification on CT (by expert annotators) and AD quantification on DRR from CT. Moreover, imaging variables add information that improves prognostic performance and is not redundant with other CVs. Using POa from CXR or POv from CT, particularly if computed by AI algorithms, has the additional advantage of simplicity, reproducibility, and speed over semiquantitative scoring systems, besides being applicable in environments that may lack access to expert human radiologists.

Our study has several limitations. Given that we selected a subset of patients who fulfilled multiple inclusion criteria (positive RT-PCR for SARS-CoV-2, and paired chest CT and CXR performed within 48 h of each other and within 48 h of presentation) and due to the single center design, we have a relatively small sample size compared with the total population of COVID-19 patients. Notwithstanding, our models generated statistically significant results and our random sample is representative of the COVID-19 patients admitted to the hospital. The AD annotation on CT was performed by two human readers, without automated algorithms; however, outlier cases were secondarily reviewed by two additional expert radiologists and corrected. While every patient in the cohort had COVID-19 at the time the CXR and CT were obtained, it is possible that not all AD detected was a manifestation of COVID-19, though expert human radiologists made their best effort to exclude likely chronic opacities on CT. Our sample size limited the number of features which could be tested at once, either through random forest or by logistic regression, which we addressed using dimension reduction techniques. Furthermore, this study used binarized clinical laboratory values—as more research establishes relative value ranges associated with severity of outcome, it would be feasible to more fairly assess the relative prognostic value of laboratorial tests compared with imaging. However, given the inconsistent panel of laboratory values obtained per patient—largely dependent on clinician discretion—such binarized laboratory values allowed for better imputation of missing values. Finally, our models' AUC of 0.75 to 0.78 suggests the existence of other predictors of COVID-19-related ICU admission not accounted for in this analysis. This performance may also suggest that different transformations of existing data—such as the use of different thresholds for laboratory values, as exist for example in other disease states—are more relevant for prognostication.

A future direction of our research includes expanding the sample size to explore more potential predictors and their interactions, as well as expanding the dataset to cover multiple timepoints for longitudinal analysis, which will likely lead to more accurate predictive models. Another direction is to include patients with CXR without CT to expand our sample size, as we have demonstrated prognostic equivalence in imaging modalities.

5 Conclusions

In conclusion, our study demonstrates that prediction of severe COVID-19 with sufficient accuracy as to be clinically relevant is attainable using only a small set of variables at initial presentation, while emphasizing that quantitative imaging variables derived from CXR and CT are superior to well established CVs without imaging, adding prognostic prediction accuracy without being redundant. Our study also emphasizes the potential of statistical prognostication models, which incorporate data from multiple modalities, including clinical, demographic, laboratory, and imaging variables. Given that AI methods can obtain quantitative imaging metrics with similar accuracy but much faster than expert human readers, AI methods for detection and quantification of AD, particularly when applied to much more widely available CXRs, feeding statistical predictive machine learning models of severe COVID-19, may augment the role of thoracic imaging and impact the management and prognostication of patients affected by COVID-19, fostering better patient outcomes, improving resource allocation and potentially increasing radiologists' efficiency.

6 Appendix: Properties of the Training and Validation Data used for Development of the AI Lung and Airspace Disease Segmentation on CT and CXR, Compared to Our Study Cohort

Table 6 contains further details on the properties of the training and validation data used for development of the AI lung and airspace disease segmentation on CT and CXR, compared to our cohort.

Table 6 Demographics, scanner, and image characteristics of data used for AI development and for the study cohort.

	AI system training CT	AI system training CXR	AI system validation CT	Study cohort CT
Datasets	Total: 1929, COVID-19: 1005, ILD: 267, pneumonia: 147, normal: 510	Total: 727 COVID-19: 0, control: 727	Total: 182, COVID-19: 86, control: 96	Total: 200, COVID-19: 200
Data origin	Multiple sites including USA, Spain, Switzerland, Germany, France, Denmark, Canada, and Belarus	Multiple sites including USA and Germany	Multiple sites including USA, Spain, and Czech Republic	US academic medical system
Sex	Female: 628, male: 827, unknown: 474	Female: 142 male: 123 unknown: 461	Female: 66, male: 101, unknown: 15	Female: 101, male: 99
Age (years)	Median: 61	Median: 58	Median: 62	Median: 59
Scanner manufacturer	GE: 450, Siemens: 1258, Philips: 41, Toshiba: 23, Other/Unknown: 156	'FUJIFILM': 111, 'Carestream': 130, 'Agfa': 25, other/unknown: 461	GE: 60, Siemens: 58, Philips: 24, Toshiba: 27, other/unknown: 13	GE: 45, Siemens: 155, Philips: 0, Toshiba: 0 other/unknown: 0
Slice thickness (mm)	$\leq$ 1.5: 1632 (1.5, 3.0]: 282, >3.0: 12	N/A	$\leq$ 1.5: 51, (1.5, 3.0]: 116, >3.0: 15	$\leq$ 1.5: 200
Reconstruction kernel	Soft: 691, hard: 1035, unknown 203	N/A	Soft: 86, hard: 71, unknown: 25	Soft: 80, hard: 120, unknown: 0

Disclosures

The senior corresponding author (Dr. Eduardo Mortani Barbosa Jr), Hae-Min Jung, Rochelle Yang, and Dr. Warren Gefter are independent authors, affiliated with the University of Pennsylvania, and while partially supported by government and industry research contracts (including from Siemens Healthineers), they have not received any compensation for any work related to this research project. Drs. Florin C. Ghesu, Boris Mailhe, Awais Mansoor, Sasa Grbic, Dorin Comaniciu, Sebastian Vogt are Siemens Healthineers employees. This research was performed under an approved IRB protocol for human data.

References

1. John Hopkins University, "COVID-19 global map," (2020). https://coronavirus.jhu.edu/map.html.
2. C. Shen et al., "Quantitative computed tomography analysis for stratifying the severity of coronavirus disease 2019," *J. Pharm. Anal.* **10**(2), 123–129 (2020).
3. K. Li et al., "CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19)," *Eur. Radiol.* **30**, 4407–4416 (2020).
4. L. Huang et al., "Serial quantitative chest CT assessment of COVID-19: a deep learning approach," *Radiol. Cardiothorac. Imaging* **2**(2), e200075 (2020).

5. T. Ozturk et al., "Automated detection of COVID-19 cases using deep neural networks with x-ray images," *Comput. Biol. Med.* **121**, 103792 (2020).
6. T. Mahmud, M. A. Rahman, and S. A. Fattah, "CovXNet: a multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization," *Comput. Biol. Med.* **122**, 103869 (2020).
7. J. P. Kanne, "Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist," *Radiology* **295**(1), 16–17 (2020).
8. D. Byrne et al., "RSNA expert consensus statement on reporting chest CT findings related to COVID-19: interobserver agreement between chest radiologists," *Can. Assoc. Radiol. J.* **72**(1), 159–166 (2021).
9. S. Zaim et al., "COVID-19 and multiorgan response," *Curr. Probl. Cardiol.* **45**(8), 100618 (2020).
10. L. Wynants et al., "Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal [published correction appears in BMJ. 2020 Jun 3;369:m2204]," *BMJ* **369**, m1328 (2020).
11. H. Swapnarekha et al., "Role of intelligent computing in COVID-19 prognosis: a state-of-the-art review," *Chaos Solitons Fractals* **138**, 109947 (2020).
12. B. T. Garibaldi et al., "Patient trajectories among persons hospitalized for COVID-19: a cohort study [published correction appears in Ann Intern Med. 2021 Jan;174(1):144]," *Ann. Intern. Med.* **174**(1), 33–41 (2021).
13. M. Francone et al., "Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis," *Eur. Radiol.* **30**(12), 6808–6817 (2020).
14. F. Ufuk et al., "The prognostic value of pneumonia severity score and pectoralis muscle area on chest CT in adult COVID-19 patients," *Eur. J. Radiol.* **131**, 109271 (2020).
15. K. Zhang et al., "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography [published correction appears in Cell. 2020 Sep 3;182(5):1360]," *Cell* **181**(6), 1423–1433.e11 (2020).
16. Z. Liu et al., "Association between initial chest CT or clinical features and clinical course in patients with coronavirus disease 2019 pneumonia," *Korean J. Radiol.* **21**(6), 736–745 (2020).
17. M. Yu et al., "Prediction of the development of pulmonary fibrosis using serial thin-section CT and clinical features in patients discharged after treatment for COVID-19 pneumonia," *Korean J. Radiol.* **21**(6), 746–755 (2020).
18. M. Wang et al., "Time-dependent changes in the clinical characteristics and prognosis of hospitalized COVID-19 patients in Wuhan, China: a retrospective study," *Clin. Chim. Acta.* **510**, 220–227 (2020).
19. M. Belgiu and L. Drăguţ, "Random forest in remote sensing: a review of applications and future directions," *ISPRS J. Photogramm. Rem. Sens.* **114**, 24–31 (2016).
20. B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genet.* **19**(Suppl. 1), 65 (2018).
21. C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.* **3**(2), 185–205 (2005).
22. M. A. Sulaiman and J. Labadin, "Feature selection based on mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1–6 (2015).
23. E. J. Mortani Barbosa, Jr. et al., "Automated detection and quantification of COVID-19 airspace disease on chest radiographs: a novel approach achieving expert radiologist-level performance using a deep convolutional neural network trained on digital reconstructed radiographs from computed tomography-derived ground truth [published online ahead of print, 2021 Jan 19]," *Invest. Radiol.* (2021).
24. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
25. O. Ronneberger, P. Fischer, and T. Brox, "U-Net convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241.

26. S. Gündel et al., "Robust classification from noisy labels: integrating additional knowledge for chest radiography abnormality assessment," *Med. Image Anal.* **72**(1), 102087 (2021).
27. N. De Jay et al., "mRMRe: an R package for parallelized mRMR ensemble feature selection," *Bioinformatics* **29**(18), 2365–2368 (2013).
28. A. Liaw and M. Wiener, "Classification and regression by random forest," *R News* **2**(3), 18–22 (2002).
29. E. LeDell, M. Petersen, and M. van der Laan, "Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates," *Electron. J. Stat.* **9**(1), 1583–1607 (2015).
30. X. Robin et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinf.* **12**, 77 (2011).

Hae-Min Jung received his BA degree in statistics from Williams College in 2017 and is currently a third-year medical student at the Perelman School of Medicine at the University of Pennsylvania.

Eduardo J. Mortani Barbosa Jr., MD, is currently an assistant professor of radiology at the University of Pennsylvania, with clinical expertise in cardiothoracic imaging and research expertise in quantitative and functional imaging utilizing machine learning and artificial intelligence methods, as well as health services research focusing on addressing disparities and creating greater value in healthcare.

Biographies of the other authors are not available.