Radiology: Artificial Intelligence

ORIGINAL RESEARCH

# Combining Initial Radiographs and Clinical Variables Improves Deep Learning Prognostication in Patients with COVID-19 from the Emergency Department

*Young Joon (Fred) Kwon, PhD, MSE • Danielle Toussie, MD • Mark Finkelstein, MD • Mario A. Cedillo, MD • Samuel Z. Maron, MA • Sayan Manna, BS • Nicholas Voutsinas, MD • Corey Eber, MD • Adam Jacobi, MD • Adam Bernheim, MD • Yogesh Sean Gupta, DO • Michael S. Chung, MD • Zahi A. Fayad, PhD • Benjamin S. Glicksberg, PhD • Eric K. Oermann, MD • Anthony B. Costa, PhD*

From the Department of Diagnostic, Molecular, and Interventional Radiology (Y.J.F.K., D.T., M.F., M.A.C., S.Z.M., S.M., N.V., C.E., A.J., A.B., Y.S.G., M.S.C., Z.A.F.), Department of Neurosurgery (Y.J.F.K., E.K.O., A.B.C.), Sinai BioDesign (Y.J.F.K., A.B.C.), BioMedical Engineering and Imaging Institute (Z.A.F.), Mount Sinai COVID Informatics Center (Z.A.F., B.S.G.), and The Hasso Plattner Institute for Digital Health at Mount Sinai (B.S.G.), Icahn School of Medicine at Mount Sinai, 1 Gustave L Levy Place, Box 1136, New York, NY 10029-6574. Received May 11, 2020; revision requested June 25; revision received November 20; accepted December 2. **Address correspondence to** Y.J.F.K. (e-mail: *fred.kwon@icahn.mssm.edu*).

Supported by the Radiological Society of North America Medical Student Research grant and the National Institutes of Health T32 Medical Scientist Training Program grant.

Conflicts of interest are listed at the end of this article.

*Radiology: Artificial Intelligence* **2021; 3(2):e200098 • https://doi.org/10.1148/ryai.2020200098 • Content codes: AI CH**

***Purpose:*** To train a deep learning classification algorithm to predict chest radiograph severity scores and clinical outcomes in patients with coronavirus disease 2019 (COVID-19).

***Materials and Methods:*** In this retrospective cohort study, patients aged 21–50 years who presented to the emergency department (ED) of a multicenter urban health system from March 10 to 26, 2020, with COVID-19 confirmation at real-time reverse-transcription polymerase chain reaction screening were identified. The initial chest radiographs, clinical variables, and outcomes, including admission, intubation, and survival, were collected within 30 days ($n$ = 338; median age, 39 years; 210 men). Two fellowship-trained cardiothoracic radiologists examined chest radiographs for opacities and assigned a clinically validated severity score. A deep learning algorithm was trained to predict outcomes on a holdout test set composed of patients with confirmed COVID-19 who presented between March 27 and 29, 2020 ($n$ = 161; median age, 60 years; 98 men) for both younger (age range, 21–50 years; $n$ = 51) and older (age >50 years, $n$ = 110) populations. Bootstrapping was used to compute CIs.

***Results:*** The model trained on the chest radiograph severity score produced the following areas under the receiver operating characteristic curves (AUCs): 0.80 (95% CI: 0.73, 0.88) for the chest radiograph severity score, 0.76 (95% CI: 0.68, 0.84) for admission, 0.66 (95% CI: 0.56, 0.75) for intubation, and 0.59 (95% CI: 0.49, 0.69) for death. The model trained on clinical variables produced an AUC of 0.64 (95% CI: 0.55, 0.73) for intubation and an AUC of 0.59 (95% CI: 0.50, 0.68) for death. Combining chest radiography and clinical variables increased the AUC of intubation and death to 0.88 (95% CI: 0.79, 0.96) and 0.82 (95% CI: 0.72, 0.91), respectively.

***Conclusion:*** The combination of imaging and clinical information improves outcome predictions.

*Supplemental material is available for this article.*



Artificial intelligence (AI) has demonstrated promise in facilitating triage in radiology departments owing to its ability to rapidly extract key features from imaging studies and perform high-throughput analysis, especially in institutions with high volumes of disease (1). Prior studies have evaluated the clinical value of AI in screening for or diagnosing coronavirus disease 2019 (COVID-19), predominantly at chest CT (2–4). In clinical practice, however, chest radiography is the primary, and often only, imaging modality used to evaluate patients with COVID-19, particularly in health systems with limited resources (5). Although the value of chest radiography in diagnosing COVID-19 might be limited by its reported low sensitivity, it may be useful in the prognostication of outcomes in patients with findings positive for COVID-19 (6–8).

Deep learning (DL) is a type of AI in which data are processed iteratively through multilayered neural networks to automatically extract high-level features from raw data input. This recursive method allows programs to discern patterns without explicit human guidance (9). Recently, a DL algorithm was reported to have accurately predicted long-term outcomes from single chest radiographs in patients with prostate, lung, colorectal, and ovarian cancer (10). Another cohort from Italy showed data that support the role of chest radiography as a first-line triage tool in predicting mild disease course of COVID-19, as defined by no need for inpatient hospitalization or inpatient hospitalization of less than 4 days duration without the need for assisted ventilation (11). A growing body of literature on the use of chest radiography shows increased severity to be associated with worse outcomes for all patients (12).

### Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, COVID-19 = coronavirus disease 2019, DL = deep learning, ED = emergency department


### Summary

Initial emergency department chest radiography and clinical variables of patients with coronavirus disease 2019 were used to train a deep learning classification algorithm to predict clinical outcomes.


### Key Points

- A deep learning algorithm trained on only routine chest radiography (intubation area under the receiver operating characteristic curve [AUC], 0.66; death AUC, 0.59) or clinical laboratory values (intubation AUC, 0.64; death AUC, 0.59) prognosticated 30-day intubation and death better than a naive classifier.
- Performance of prediction of intubation (AUC, 0.88) and death (AUC, 0.82) increased when the model was trained with initial chest radiographs and relevant clinical variables from electronic health records acquired exclusively from the emergency department encounter.
- The model, despite training with only young patients aged 21–50 years, can be generalized to a pseudoprospective test set that also included patients older than 50 years.

Some DL algorithms incorporating CT and chest radiography data have been used to aid in screening for and diagnosis of COVID-19, and one study used CT to predict poor prognostic outcomes in patients with COVID-19 (2,13–15). A model that predicts a pulmonary x-ray severity score based on chest radiographs from patients with COVID-19 was published in 2020 (16). Nonetheless, the potential for DL algorithms (especially those that have been trained using chest radiographs from patients with COVID-19 and clinically validated severity scores provided by expert radiologists) to directly predict clinical outcomes and to aid in prognostication and risk stratification based on only the chest radiograph as input has gone largely unexplored (8). In fact, many recently published prognostication algorithms use only clinical variables and either do not use imaging data as input or use only CT images, the latter of which is problematic because CT is less widely available and less frequently performed than chest radiography (17–19).

The presence of comorbidities, such as lung and heart disease, can potentially confound interpretation of chest radiographs in patients with COVID-19 pneumonia, which may decrease the predictive ability of DL (20). Thus, in this context, the generation of predictive chest radiograph interpretations may be more valid in patients younger than 50 years, who have a lower prevalence of such comorbidities. While COVID-19 affects persons of all ages, the younger population comprises a considerable proportion of affected patients (21). Thus, testing for generalizability of prognostication algorithms in patients with COVID-19 is important for deployment of DL to appropriate patient populations.

In this study, we propose a proof-of-concept model with intent to demonstrate that a DL algorithm can take only the initial chest radiograph—an imaging study that emergency department (ED) clinicians do not routinely use as the main determinant of hospitalization—and the clinical variables from the ED to prognosticate the outcomes of patients with COVID-19 (8). We compared the performance of the model trained on chest radiographs or clinical variables alone with that of the model trained on both chest radiographs and clinical variables to evaluate the individual contribution of chest radiographs or clinical variables to the prognostication and to test for a potential synergistic effect of combining the two types of inputs. To do so, we used a DL classification algorithm previously used to predict 14 different diseases, including pneumonia, based on chest radiographs (22). We hypothesized that training the convolutional neural network with image input and the associated chest radiograph severity score, which was previously reported and validated in Toussie et al (8), is as effective as training with the image input and the associated clinical outcome of admission as labels. We then tested this model to generate a model severity score that was distinct from the expert radiologist–generated severity score, using only the image data input on an unseen test set of patients of all ages (including patients older than 50 years) who presented at different time points to predict hospital admission, need for intubation, and risk of mortality. To improve the model performance, we also supplemented the model with standard laboratory tests available at the initial ED encounter.

## Materials and Methods

### Patient Selection

To collect the patient cohort for this institutional review board–approved retrospective cohort study in which written consent was waived, we used the Montage Search and Analytics platform (Montage Healthcare Solutions) and extracted radiology information system data from all chest radiographic examinations performed in the ED setting from March 10 to 29, 2020, in three hospitals in New York City with different radiography acquisition devices (Table 1). We removed any protected health information from the patient data for analysis and ensured the study was compliant with the Health Insurance Portability and Accountability Act. We used the obtained cohort to extract relevant clinical and laboratory data from the electronic medical record. The resulting radiology information system dataset contained 4738 ED encounters. The exclusion criteria included age older than 50 years or younger than 21 years ($n$ = 3163), duplicate chest radiograph in the same patient ($n$ = 81), patients with an unconfirmed COVID-19 real-time reverse-transcription polymerase chain reaction positive test result ($n$ = 1101), presentations unrelated to COVID-19 ($n$ = 2), unevaluable chest radiograph ($n$ = 1), and inaccessible clinical data ($n$ = 1). All 338 patients from the original Toussie et al (8) clinical study were included in the training and validation dataset. We used the data to train a prognostication DL algorithm, a different purpose and outcome assessment from those of the original clinical study, in which expert radiologists scored the chest radiographs directly.

We randomly assigned the included chest radiographs obtained between March 10 and March 26, 2020, ($n$ = 338) to

**Table 1: Patient Characteristics from the Training, Validation, and Test Datasets**

| Variable | Overall | Training | Validation | Test | *P* Value |
|---|---|---|---|---|---|
| Total | 499 | 283 | 55 | 161 | NA |
| Men | 308 (62) | 174 (62) | 36 (65) | 98 (61) | .83 |
| Age (y) | 42 (34–50) | 38 (31–45) | 41 (35–44) | 60 (46–70) | <.001 |
| Race | | | | | .1 |
| White | 99 (20) | 59 (21) | 12 (22) | 28 (17) | … |
| Asian | 43 (9) | 28 (10) | 2 (4) | 13 (8) | … |
| Black | 114 (23) | 65 (23) | 13 (24) | 36 (22) | … |
| Hispanic | 161 (32) | 95 (34) | 21 (38) | 45 (28) | … |
| Other or unknown | 82 (16) | 36 (13)* | 7 (13)* | 39 (24)* | … |
| BMI | 29 (24–36) | 29 (24–36) | 32 (26–39) | 28 (24–32) | .01 |
| BMI cutoffs | | | | | .01 |
| Normal | 135 (29) | 81 (28) | 11 (20) | 43 (33) | … |
| Overweight | 126 (27) | 78 (27) | 11 (20) | 37 (28) | … |
| Mild or moderate obesity | 138 (28) | 74 (26) | 20 (36) | 44 (33) | … |
| Severe obesity | 71 (14) | 50 (18)* | 13 (24) | 8 (6) | … |
| Smoker | | | | | .15 |
| No | 327 (66) | 186 (66) | 37 (67) | 104 (65) | … |
| Former | 58 (12) | 25 (9) | 5 (9) | 28 (17) | … |
| Other or unknown | 86 (17) | 54 (19) | 10 (18) | 22 (14) | … |
| Yes | 28 (6)* | 18 (6) | 3 (5)* | 7 (4) | … |
| Site | | | | | .39 |
| Manhattan | 201 (40) | 118 (42) | 25 (45) | 58 (36) | … |
| Brooklyn | 154 (31) | 90 (32) | 12 (22) | 52 (32) | … |
| Queens | 144 (29) | 75 (27)* | 18 (33) | 51 (32) | … |

Note.—Continuous variables shown as mean, with interquartile range in parentheses. Categorical variables are shown as number of patients, with the percentage in parentheses. BMI = body mass index, NA = not applicable.
* Percentages do not equal 100 due to rounding.

either the training set (*n* = 283 [84%]) or the validation set (*n* = 55 [16%]) for the DL model. All radiographs were obtained with bedside imagers. In the training set, 73.5% (208 of 283) were anteroposterior radiographs and 26.5% (75 of 283) were posteroanterior and lateral radiographs. In the validation set, 76.4% (42 of 55) were anteroposterior radiographs, and 23.6% (13 of 55) were posteroanterior and lateral radiographs. We used only frontal radiographs for model training. The included chest radiographs from March 27 to March 29, 2020, (*n* = 51) were assigned to compile a held-out test set from a different time (Fig 1). A total of 161 patients were included in the test set. Of these patients, 51 were between the ages of 21 and 50 years, and 110 were older than 50 years. These 110 patients were added to test for generalizability of the model in older patients at increased risk of mortality. Twenty-nine patients in the test set did not have sufficient information (missing either height and/or weight) to calculate body mass index. In the test set of patients aged 21 to 50 years (*n* = 51), 68.6% (35 of 51) of the radiographs were anteroposterior views and 31.4% (16 of 51) were posteroanterior and lateral views. In the test set of patients older than 50 years (*n* = 110), 96 radiographs (87.3%) were anteroposterior views and 14 (12.7%) were posteroanterior and lateral views (Table 2). We used only frontal radiographs for model inference.

## Data Collection

Two fellowship-trained cardiothoracic radiologists (C.E., 26 years of experience; Y.S.G., 1 year of experience) blinded to patient history other than COVID-19 positivity independently examined the initial chest radiographs for opacities to generate a total severity score. Each lung was divided into three zones (upper, middle, lower), and a binary score of 0 (no opacity) or 1 (opacity) was assigned to each lung zone (Fig E1 [supplement]) (8). For model training, only the lung zones that both radiologists agreed contained opacity were given the final opacity label (score of 1); otherwise, the lung zones were deemed normal (score of 0). Chest radiographs with a score of 2 or higher (out of 6) were categorized as severe for the purposes of the training algorithm. Any admission, intubation, or death during 30-day follow-up was categorized as a positive event.

## Model Architecture and Training

We stripped the raw images of any metadata for de-identification. We resized and center cropped the radiographs to a

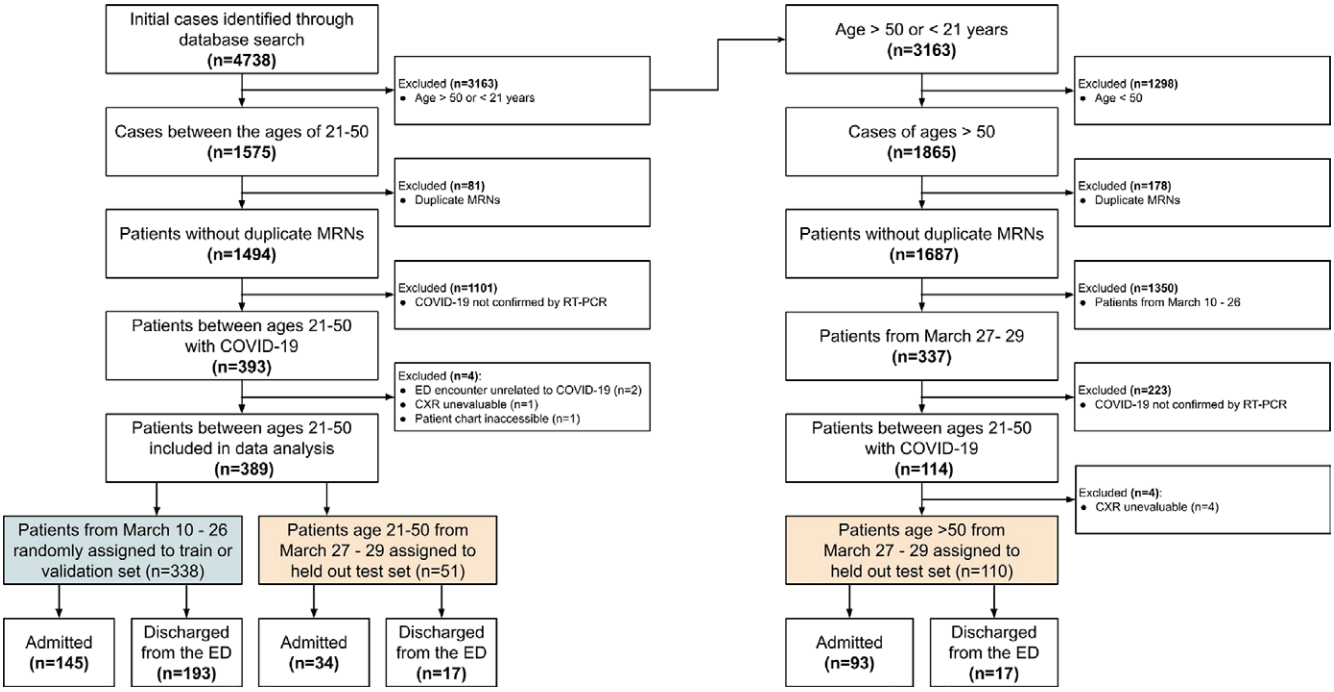



**Figure 1:** Patient inclusion and exclusion criteria. CXR = chest radiograph, COVID-19 = coronavirus disease 2019, ED = emergency department, MRN = medical records number, RT-PCR = reverse-transcription polymerase chain reaction.

**Table 2: Distribution of Imaging Modality, Severe Chest Radiography, and Clinical Outcomes**

| Variable | Total (*n* = 499) | Training (*n* = 283) | Validation (*n* = 55) | Test (*n* = 161) |
|---|---|---|---|---|
| Modality | | | | |
| Bedside | 381 (76.4) | 208 (73.5) | 42 (76.4) | 131 (81.4) |
| PA and lateral | 118 (23.6) | 75 (26.5) | 13 (23.6) | 30 (18.6) |
| Chest radiograph severity score | 248 (49.7) | 111 (39.2) | 27 (49.1) | 110 (68.3) |
| 30-day admission | 271 (54.3) | 121 (42.8) | 27 (49.1) | 123 (76.4) |
| 30-day intubation | 73* (14.8) | 20 (7.1) | 7 (12.7) | 46* (29.5) |
| 30-day mortality | 51 (10.2) | 8 (2.8) | 2 (3.6) | 41 (25.5) |

Note.—Data are numbers of patients, with percentages in parentheses. Only frontal views from posteroanterior (PA) and lateral acquisitions were used for training. The test set includes 110 patients older than 50 years.
* The 30-day intubation value in the test set excludes five patients older than 50 years who were indicated as "do not intubate."

resolution of 1024 × 1024. The authors visually inspected all radiographs after cropping, which standardized input size and removed any texts that were embedded in the edges of some radiographs (eg, time of acquisition). The images were subsequently converted to tensors and normalized with the ImageNet *(http://image-net.org)* mean and standard deviation. They were stored as HDF5 (The HDF Group) datasets to prevent the need to preprocess the images for each iteration of training. For the prediction algorithm, we used the DenseNet-121 architecture that was first pretrained on ImageNet, a model previously used in the CheXNet study (22–24). We used two different labeling schemes for the training*: (a)* radiographs with the associated expert-generated severity scores as labels or *(b)* radiographs with the associated admission status as labels as a control. The DenseNet-121 output was then compiled by a fully connected layer and a sigmoid function to generate a probability score for the label (ie, severe, not severe or admitted, or not admitted). We used the binary cross-entropy loss function and the Adam optimizer (25) (Fig 2). We empirically tested for the best learning rate from $1 \times 10^{-2}$ to $1 \times 10^{-10}$ in logarithmic increments ($1 \times 10^{-2}$, $1 \times 10^{-3}$, …, $1 \times 10^{-10}$) and determined the best learning rate as one that resulted in the lowest validation loss after 10 epochs of training.

We also tested how the model performance would change with the addition of the following clinical variables initially acquired in the ED from electronic health records: white blood cell count; C-reactive protein, D-dimer, lactate, lactate dehydrogenase, creatinine, troponin, aspartate aminotransferase,

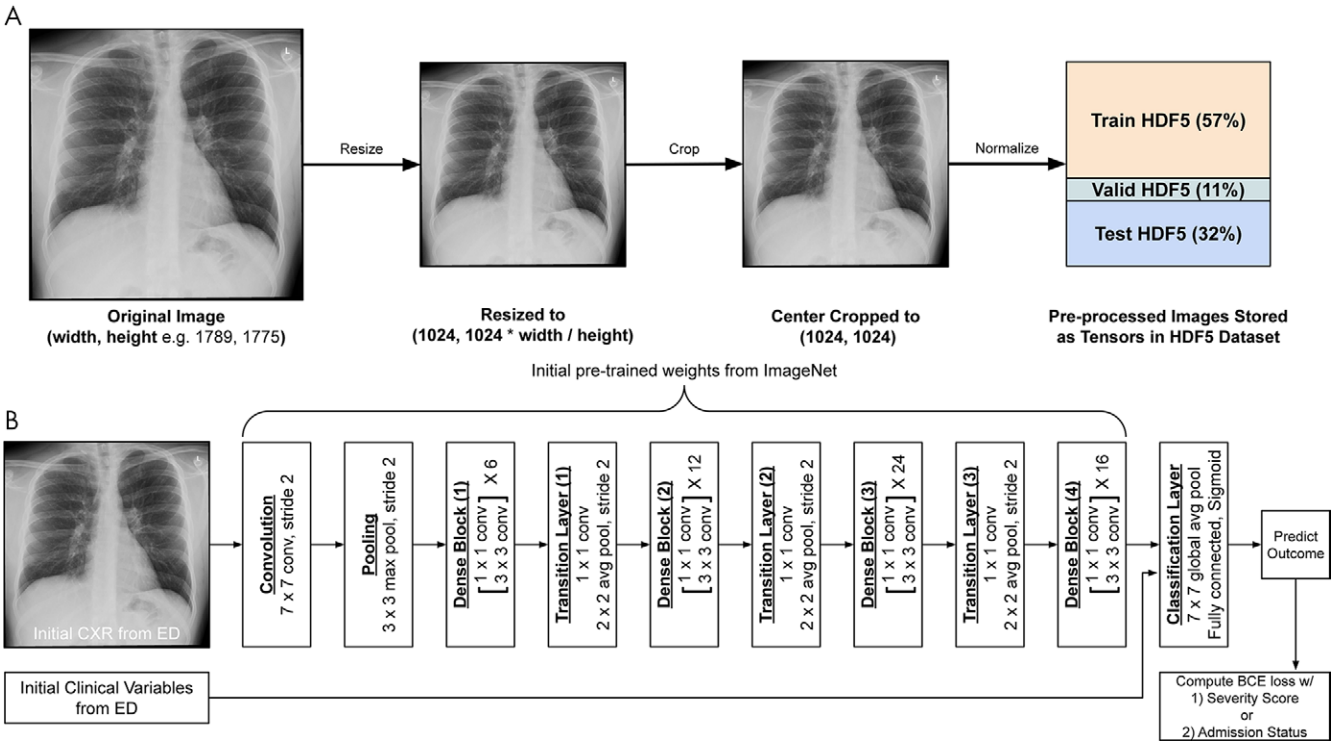



**Figure 2:** *A*, Preprocessing of radiographs and storage as HDF5 (The HDF Group) datasets. When images are stored as HDF5 datasets, they do not require preprocessing (eg, resizing, cropping, conversion to tensors) each time they are loaded to memory. *B*, Model architecture and training scheme. The two training methods we conducted included computing binary cross-entropy (BCE) loss function with either severity score (1 for severe, 0 for not severe) or admission status (1 for admitted, 0 for not admitted within 30 days). For inference, the deep learning algorithm outputs a severity score (distinct from radiologist-generated severity score) based on the chest radiograph (CXR) alone that is used to predict admission. To better predict intubation and death, initial clinical variables from the emergency department (ED) were added and used to retrain a model previously trained on chest radiographs and severity score. Avg = average, conv = convolutions.

and glucose levels; estimated glomerular filtration rate; and systolic and diastolic blood pressure. We used mean imputation for any unavailable laboratory values. For model training with clinical variables alone, we used fully connected layers. For model training with both the chest radiograph and the clinical variables, the clinical variables were concatenated and added as input before the fully connected layer in the classification layer of the DenseNet-121 model previously trained on the chest radiograph and its previously mentioned severity score (Fig 2).

## Model Evaluation

We selected the model from the training with either the chest radiograph severity score or the admission status with the minimum validation set loss as the best model to test. The probability score output, a continuous floating point value from 0 to 1 and distinct from the ordinal integer grading score from expert radiologists, from the DL algorithm based on only the chest radiograph as input was used to calculate the area under the receiver operating characteristic curve (AUC) for four different classes: chest radiograph severity scores, admissions, intubations, and deaths. For example, a DL algorithm–generated score greater than 0.65 predicts admission, a score greater than 0.80 predicts intubation, and a score greater than 0.90 predicts death. To account for variable prevalence among classes, we designated classes with a prevalence greater than or equal to 40% in our cohort as the majority class, while those with a prevalence less than 40% were designated the minority class. Severe chest radiographs and admissions were thereby majority classes, while intubation and death were minority classes. We then plotted the precision-recall curve to evaluate the model performance for minority classes that were not used as part of the training. We used the discriminative localization methods previously described to generate heatmaps that describe which areas of the radiographs were contributing the most to the prediction algorithm (26,27). The source code used in this article is publicly available at *https://github.com/aisinai/covid19_cxr*.

## Statistical Analysis

Bivariate analysis of continuous variables, such as body mass index and age, was performed using the Kruskal-Wallis *H* test. Bivariate analysis of categorical variables, such as race, sex, smoking history, hospital site, and comorbidities, was performed with the $\chi^2$ test.

To calculate the AUC, accuracy, precision, recall, and F1 score values, an operating point was selected for high sensitivity (recall), which was then used to calculate accuracy and F1 score. To calculate 95% CIs for AUC, accuracy, precision, recall, and F1 score values, we used bootstrapping experiments, as previously described (28–30). We resampled the test set with replacement and repeated the inference 100 000 times. The resampled test set was the same size as the original test set

($n$ = 161) because we were approximating the variation of the statistic that depends on the sample size. We compared the computed statistics with those of a naive classifier that predicts the positive class every time (ie, the naive model always predicts severe chest radiograph, 30-day admission, intubation, and death).

## Results

### Patient Demographics

Overall, 499 patients and their chest radiographs were included between the training, validation, and test sets, with a diverse patient population. Of the 499 chest radiographs scored, 41 (8.2%) had a severity score of 2 or higher given by one of the two reviewers, but not when the severity score was calculated with concordant scores. The remaining 458 chest radiographs (91.8%) had been correctly categorized as severe (score of ≥2) or not severe (score of 0 or 1) by both reviewers individually and by concordant scores. Of the 499 patients (median age, 42 years [interquartile range, 34–50]; 308 men), 248 (49.7%) had severe chest radiographs, 271 (54.3%) were admitted, 73 (14.8%) were intubated (five patients had "do not intubate" status), and 51 (10.2%) died. Additionally, there were 53 patients (10.6%) with asthma, three (0.6%) with chronic obstructive pulmonary disease, 105 (21.0%) with hypertension, 73 (14.6%) with diabetes mellitus, seven (1.4%) with HIV, 18 (3.6%) with cancer, 23 (4.6%) with chronic kidney disease, 25 (5.0%) with coronary artery disease, and five (1%) with atrial fibrillation. The datasets differed significantly with regard to age (due to inclusion of patients younger than 50 years in the test set) and body mass index ($P$ = .01) (Table 1). Otherwise, there were no significant differences in the distribution of demographic information between the training, validation, and test sets. For patients who were intubated, the time from initial chest radiography to intubation averaged 3.7 days, with a median of 3 days (range, 0–12 days).

The training, validation, and test datasets consisted of 283, 55, and 161 patients, respectively. Severe chest radiograph, admission, intubation, and death data for these datasets are found in Table 2. The subset of the test set of 51 patients aged 21–50 years had 34 (66.7%) severe chest radiographs, 34 (66.7%) admissions, 10 (19.6%) intubations, and seven (13.7%) deaths (Table 2). Of the 499 chest radiographs that were scored, 41 (8.2%) had a severity score of 2 or higher given by one of the two reviewers, but not when the severity score was calculated with concordant scores. The remaining 458 chest radiographs (91.8%) were correctly categorized as severe (score of ≥2) or not severe (score of 0 or 1) by both reviewers individually and by concordant scores.

### Model Training

Empirical search and determination of the best learning hyperparameters showed that the validation loss was lowest with the learning rate of $1 \times 10^{-5}$, β1 decay of 0.99, β2 decay of 0.9999, and weight decay of $1 \times 10^{-5}$ after 10 epochs of training. Training with the chest radiograph severity scores and the admission status converged to the best model, as evaluated by validation loss (Fig E2 [supplement]). Initially, iterations of training showed low AUC for predicting death in the validation set, but the AUC increased with additional iterations (Fig E2 [supplement]).

### Prediction of Independent Clinical Outcome Variables

After selection of the best model based on the minimum validation loss (Fig E2 [supplement]), we used the held-out previously unseen test set to produce prediction outputs. The single prediction output from each of the two models, the model trained with chest radiograph severity scores or the model trained with admissions, was then used to generate AUC values for the chest radiograph severity scores and the three clinical variables: any admission, intubation, or death event in 30 days. Both models gave satisfactory AUCs. The model trained on the chest radiograph severity score produced the following AUCs: 0.80 (95% CI: 0.73, 0.88) for chest radiograph severity score, 0.76 (95% CI: 0.68, 0.84) for admission, 0.66 (95% CI: 0.56, 0.75) for intubation, and 0.59 (95% CI: 0.49, 0.69) for death (Fig 3). Notably, the lower bound of the 95% CI for 30-day intubation prediction (0.56, 0.75) was greater than 0.5, the expected performance of a classifier without discriminative abilities. The model trained on admission status produced the following AUCs: 0.70 for chest radiograph severity score, 0.70 for admission, 0.58 for intubation, and 0.50 for death. These AUCs did not significantly differ when trained with radiographs and severity scores as labels or 30-day admission status as labels (Fig 3).

The precision versus recall (positive predictive value vs sensitivity) curves suggest that the performance was better on majority classes (chest radiograph severity score and admission status) than on minority classes (intubation status and death). The accuracy of predicting 30-day intubation status (47% [95% CI: 39, 54]) and death (42% [95% CI: 34, 50]) was nonetheless better than a naive classifier that always predicts the positive class (accuracy of 30% and 26% for 30-day intubation and death, respectively) (Table 3) (Fig 4). Further, compared with a naive classifier, the performance of negative predictive value and specificity (ie, precision for the minority class) was better at predicting lack of intubation or survival (Fig E3 [supplement]).

The model performance on the prediction of intubation and death increased when trained with clinical variables from electronic health records and with intubation status as the target label (Fig 5). The AUC increased from 0.66 to 0.88 (95% CI: 0.79, 0.96) for intubation and from 0.59 to 0.82 (95% CI: 0.72, 0.91) for death for the aggregate test dataset with all adults older than 21 years. The combined model performed better than the model trained on clinical variables alone. As expected, the model performed better for the young adults aged 21–50 years, but it still demonstrated clinically useful results for the patients older than 50 years in the test set. At our selected operating point for intubation that prioritizes recall or sensitivity greater than 80%, the F1 score remained high (>65%).

The heatmap results indicate that the inferior left side of the patient's chest (right side of the anteroposterior radiograph) that contains the heart and the gastric bubble contributes less to the

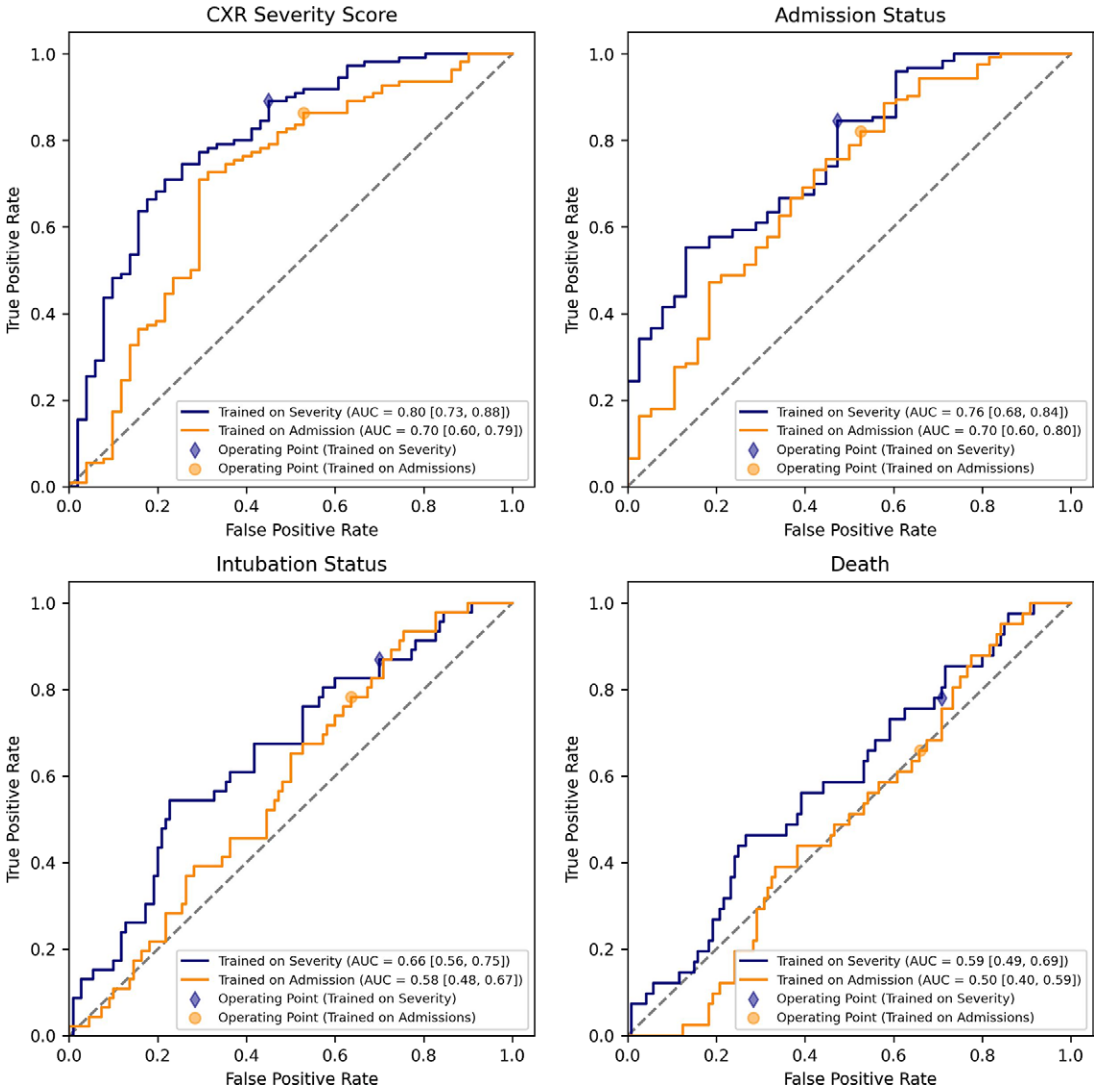



**Figure 3:** Receiver operating characteristic (ROC) curves of the test set based on two training schemes. Areas under the ROC curves (AUCs) do not differ between training schemes in terms of severity score or admission status. The 95% CIs for the AUCs are in brackets. Operating point was selected for high sensitivity (recall), which was then used for accuracy and F1 score calculations. CXR = chest radiography.

radiograph compared with the rest of the radiograph. The absolute value of the model output (given as a probability) increases with worse clinical outcomes (Fig 6).

## Discussion

We hypothesized that a DL model could predict prognosis of adult patients with COVID-19 based solely on routinely available imaging (chest radiographs) and laboratory studies in the ED. We initially selected a younger patient cohort to reduce the potential presence of comorbidities that could decrease the predictive ability of our DL algorithm (20). We then included additional patients older than 50 years in the test set to assess for generalizability of the model in older patients at higher risk. Using a previously successful DL classification algorithm, DenseNet-121, we trained the model successfully with the chest radiograph and the associated severity score or the 30-day admission status. This trained model could then take unseen chest radiographs from another time to predict 30-day admission status, intubation status, and survival, despite the differences in patient age and outcomes in the test set compared with those in the training and validation sets. We also trained a model with clinical variables alone and compared the models trained on either chest radiographs or clinical variables only to a model trained with both chest radiographs and clinical variables. The combined model had the best performance.

Fine et al (31) surveyed ED physicians as to what factors guide their decisions on whether to admit or discharge a patient with community-acquired pneumonia and found that chest radiography is not a major factor in the decision-making process. Furthermore, CURB-65 (confusion, uremia, respiratory rate, blood pressure, age ≥65 years) and the Pneumonia Severity Index—the most widely used scoring systems to guide decisions on admitting patients with community-acquired pneumonia—exclude chest radiographs as major or minor criteria (32). However, Toussie et al (8) demonstrated that the severity of opacities on chest radiographs does predict outcomes in COVID-19 pneumonia. The severity of opacity on the presentation chest radiograph is an important objective assessment of disease severity

**Table 3: Accuracy, Precision (Positive Predictive Value), Recall (Sensitivity), and the F1 Score for the Test Set as an Aggregate and as Subgroups for Patients Aged 21–50 Years or Older than 50 Years**

| | All Patients ($n$ = 161) | | | Patients Aged 21–50 Years ($n$ = 51) | | | Patients Older than 50 Years ($n$ = 110) | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Naive Classifier | Trained on Scores | Trained on Admissions | Naive Classifier | Trained on Scores | Trained on Admissions | Naive Classifier | Trained on Scores | Trained on Admissions |
| Accuracy | | | | | | | | | |
| Severity score | 68 | 78 (70, 83) | 73 (66, 80) | 71 | 86 (81, 91) | 78 (72, 84) | 67 | 74 (66, 80) | 71 (64, 78) |
| Admission status | 76 | 77 (70, 83) | 74 (67, 81) | 67 | 90 (86, 94) | 78 (72, 84) | 81 | 66 (59, 73) | 72 (65, 79) |
| Intubation status | 30 | 47 (39, 54) | 49 (41, 57) | 20 | 47 (39, 55) | 49 (41, 57) | 34 | 47 (39, 54) | 49 (41, 56) |
| Death | 26 | 42 (34, 50) | 42 (34, 49) | 14 | 45 (37, 53) | 47 (40, 55) | 31 | 40 (32, 48) | 39 (32, 47) |
| Precision | | | | | | | | | |
| Severity score | 68 | 80 (73, 87) | 78 (70, 85) | 71 | 91 (86, 96) | 86 (79, 92) | 67 | 76 (68, 83) | 74 (64, 78) |
| Admission status | 76 | 85 (79, 91) | 83 (77, 90) | 67 | 91 (86, 96) | 83 (75, 90) | 81 | 82 (75, 88) | 84 (77, 90) |
| Intubation status | 30 | 34 (26, 43) | 34 (25, 43) | 20 | 26 (18, 34) | 25 (17, 34) | 34 | 38 (29, 46) | 38 (29, 47) |
| Death | 26 | 27 (19, 35) | 25 (17, 34) | 14 | 20 (13, 28) | 19 (12, 27) | 31 | 30 (22, 39) | 28 (20, 37) |
| Recall | | | | | | | | | |
| Severity score | 100 | 89 (83, 95) | 85 (79, 92) | 100 | 89 (83, 94) | 83 (76, 90) | 100 | 84 (77, 90) | 86 (80, 93) |
| Admission status | 100 | 85 (78, 91) | 82 (75, 89) | 100 | 94 (89, 98) | 85 (78, 92) | 100 | 75 (68, 83) | 81 (74, 87) |
| Intubation status | 100 | 87 (77, 96) | 78 (66, 90) | 100 | 90 (78, 100) | 80 (65, 93) | 100 | 86 (76, 94) | 78 (66, 89) |
| Death | 100 | 78 (65, 90) | 66 (51, 80) | 100 | 100 (100, 100) | 86 (70, 100) | 100 | 74 (61, 85) | 62 (48, 75) |
| F1 score | | | | | | | | | |
| Severity score | 81 | 84 (79, 89) | 81 (76, 87) | 83 | 90 (86, 94) | 84 (79, 89) | 80 | 79 (73, 85) | 80 (74, 85) |
| Admission status | 86 | 85 (80, 89) | 83 (77, 88) | 80 | 93 (89, 96) | 84 (78, 89) | 90 | 78 (73, 84) | 82 (77, 87) |
| Intubation status | 46 | 49 (39, 58) | 47 (37, 57) | 33 | 40 (29, 50) | 38 (27, 49) | 51 | 53 (43, 61) | 51 (41, 60) |
| Death | 41 | 41 (31, 50) | 36 (26, 46) | 25 | 33 (23, 43) | 31 (20, 41) | 47 | 43 (33, 52) | 39 (29, 48) |

Note.—All values are percentages. Data in parentheses are 95% CIs.

that can be used to guide physicians in deciding if a patient needs to be admitted or can be discharged safely. We used chest radiographs and the associated scores provided by expert radiologists to train a model that requires only the initial radiograph to predict clinical outcomes for test populations of patients with COVID-19. Given that COVID-19 is still rapidly spreading across the United States and overwhelming hospitals, a quick tool that can provide accurate prognostication for COVID-19 and can help appropriately allocate resources (eg, inpatient hospital beds, ventilators) for subsequent management is vital.

The advantage of this study design lies in the ability for a DL model to predict clinical outcomes rather than to only screen for or confirm a diagnosis of COVID-19, as seen in other studies (5,7,33–36). The radiographs in the training and testing sets come from multiple hospitals across three boroughs of New York City, all with different acquisition devices. The diversity of the chest radiographs used in the model and the different time frame of the test cohort (ie, a pseudoprospective trial) suggest an increased likelihood of generalizability. While surveys of ED physicians do not typically report chest radiography findings as a major factor in the decision-making process to admit a patient with community-acquired pneumonia, this algorithm can reliably predict 30-day admission status in patients with COVID-19 and may serve as a first-pass triaging process to alert radiologists and clinicians of patients at higher risk who will likely require hospitalization (31). This prioritization of care can be readily adopted within existing clinical workflows and can lead to validation in actual clinical practice, thereby addressing the common challenges and criticisms of existing AI research in medicine (37,38).

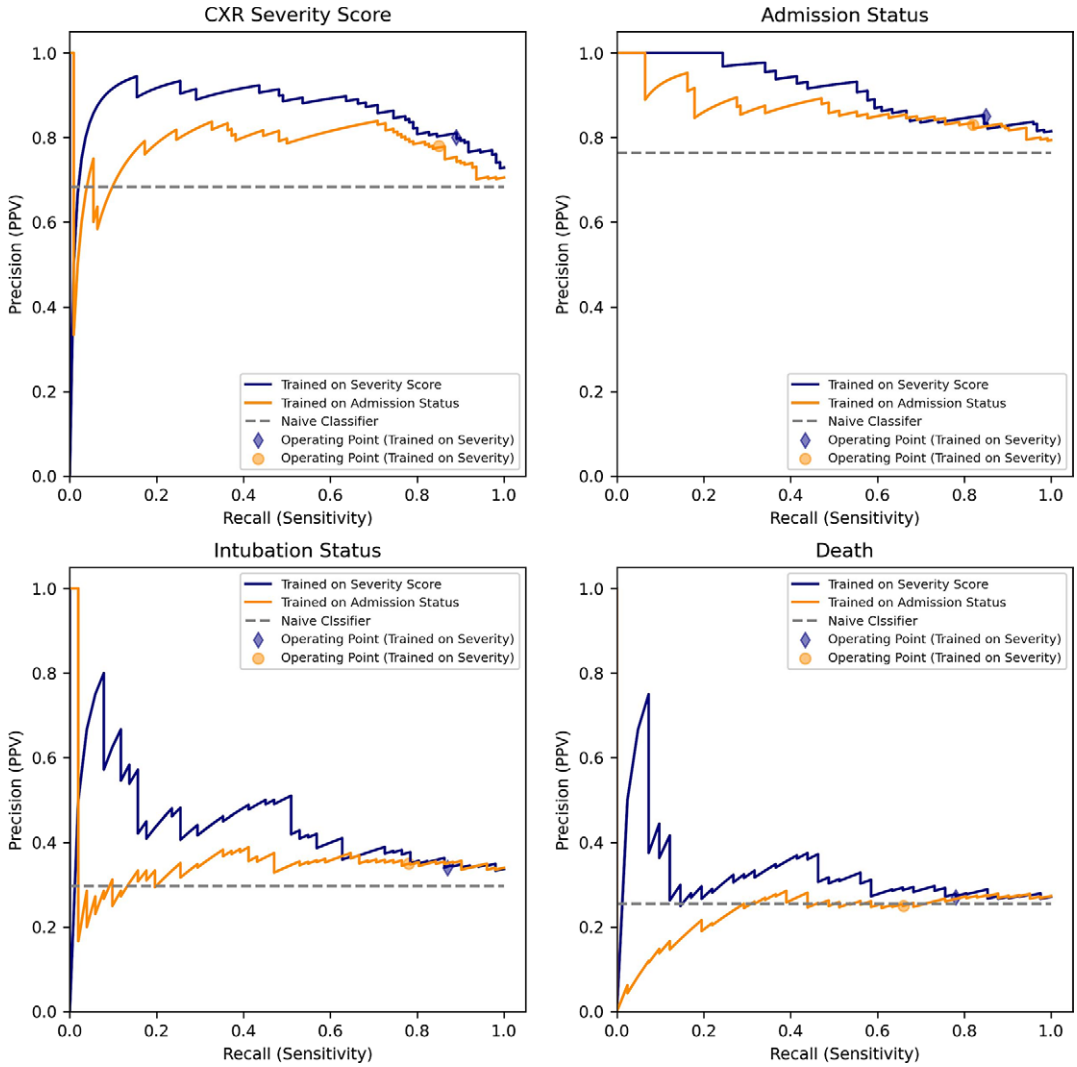
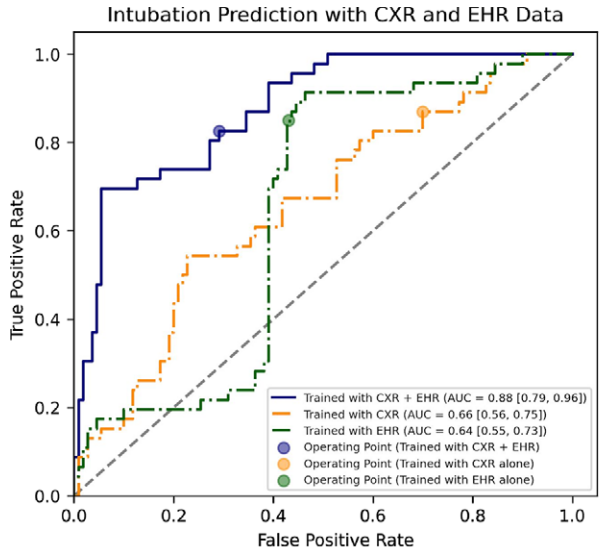


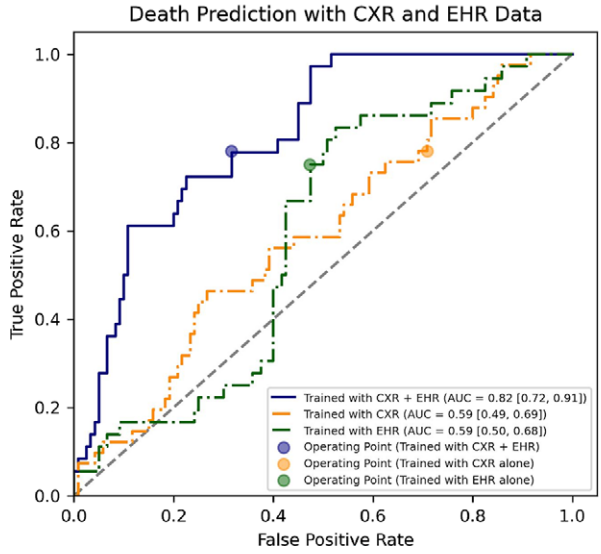

**Figure 4:** Precision versus recall curves for four prediction categories. Chest radiography (CXR) severity score and admission status were used for training and were well balanced. Intubation status and mortality were minority classes and were not seen during training, thus producing poor precision (positive predictive value [PPV]) and recall (sensitivity) performance. Nonetheless, both intubation and death predictions of this model performed better than a naive classifier that would predict a positive class each time.

This study confirms that the chest radiograph severity score can be used to train a network that predicts clinical outcomes, including the need for hospitalization, intubation, and mortality. There are multiple benefits of using the initial chest radiograph severity score rather than the outcome of interest in the prediction model. The model after deployment requires only the initial chest radiograph to provide prognostic predictions, without the need for any additional manual scoring inputs or clinical variables. From a developer's perspective, training a model using 30-day outcomes relies on the ability to follow all patients for 30 days, and some patients may be lost to follow-up after the initial ED encounter. Since the severity score is assigned to all initial chest radiographs from the ED, there is no need for 30-day follow-up. Additionally, the chest radiograph severity scores can be incorporated into the model without the need to wait 30 days to confirm the absence of an admission event. It is possible that the patients are admitted for other nonrespiratory reasons, whereas the severity score that indicates opacity in lung zones at chest radiography more directly correlates to potential intubation. Therefore, the dataset can be expanded immediately with the widespread availability of chest radiographs and the initial laboratory values from the ED. Most importantly, the algorithm outputs similar AUCs for the severity score and clinical outcomes (30-day admission, intubation, and death) in unseen test radiographs whether the algorithm is trained with either the severity scores or the clinical outcomes.

The precision and recall of the predictions, based on chest radiographs alone, for intubation and death showed lower performance than those for chest radiograph severity scores and admission because of the relative scarcity of these events. Nonetheless, this model still performed better than a naive classifier and had a better prediction performance for the negative class (ie, better negative predictive values and specificity). Our study confirms the results of Toussie et al (8) that the findings from the initial chest radiograph obtained in the ED contains information that enables physicians to better predict the need for hospitalization and help




| Intubation | All Patients (n=156) | Patients Aged 21-50 (n=51) | Patients Aged > 50 (n=105) |
|---|---|---|---|
| Accuracy | 0.74 [0.64, 0.84] | 0.98 [0.90, 1.00] | 0.63 [0.55, 0.72] |
| Precision | 0.54 [0.45, 0.64] | 1.00 [0.93, 1.00] | 0.48 [0.40, 0.57] |
| Recall | 0.82 [0.72, 0.91] | 0.90 [0.82, 0.98] | 0.81 [0.72, 0.90] |
| F1-Score | 0.66 [0.58, 0.75] | 0.95 [0.87, 1.00] | 0.60 [0.51, 0.69] |




| Death | All Patients (n=161) | Patients Aged 21-50 (n=51) | Patients Aged > 50 (n=110) |
|---|---|---|---|
| Accuracy | 0.71 [0.61, 0.81] | 0.96 [0.87, 1.00] | 0.59 [0.50, 0.69] |
| Precision | 0.42 [0.32, 0.52] | 0.78 [0.69, 0.88] | 0.37 [0.29, 0.45] |
| Recall | 0.78 [0.68, 0.87] | 1.00 [0.93, 1.00] | 0.72 [0.63, 0.81] |
| F1-Score | 0.55 [0.46, 0.65] | 0.88 [0.79, 0.98] | 0.49 [0.40, 0.58] |

**Figure 5:** The area under the receiver operating characteristic curve (AUC) of intubation prediction from a model that incorporates clinical variables from electronic health records (EHR) to the model previously trained on chest radiographs (CXR) alone and their severity score. The AUC for predicting intubation increased from 0.66 to 0.88, and the AUC for predicting death increased from 0.59 to 0.82. At the authors' selected operating point, the sensitivity remained high, while achieving a good F1 score. Intervals indicate 95% CIs. Five patients classified as "do not intubate" in the test set were excluded from intubation data analysis.

ensure that the appropriate patients are admitted or discharged. The model trained only on chest radiography had AUCs for intubation and death prediction similar to those of the model trained only with clinical variables first obtained in the ED. The model trained only with clinical variables had a low true-positive rate and a high false-positive rate at high cutoffs (left side of receiver operating characteristic curve [Fig 5]). That is, using clinical variables alone that may be limited in availability within days of the initial ED encounter, the model cannot sufficiently separate patients who require intubation at high cutoff thresholds. We improved the performance of prediction of intubation and death when both inputs and the same respective architectures to extract information were combined into one model. Our model, which uses only the information from the initial ED encounter from standard imaging and routinely ordered laboratory tests, may be used to help guide hospitalization decisions in patients with COVID-19 and inform ED physicians of the risks of their patients experiencing poor outcomes later in the disease course. The timeline to prognostication is clinically relevant, given that in this patient cohort that required intubation, the time from the first chest radiograph to intubation was a median of 3 days.

The difficulty in understanding logical reasoning of DL algorithms, especially those that predict prognostics, is an inherent challenge known as the black box problem (37–39). We used heatmaps to ensure that appropriate regions of the radiographs were contributing to the final prediction output. The heatmaps suggested that the irrelevant areas of the radiograph were not contributing substantially to the final output. Of note, heatmaps do not indicate which anatomic regions and qualities are truly contributing to the prediction. Further, the regions generated by the heatmaps may be of different sizes than the actual subregion

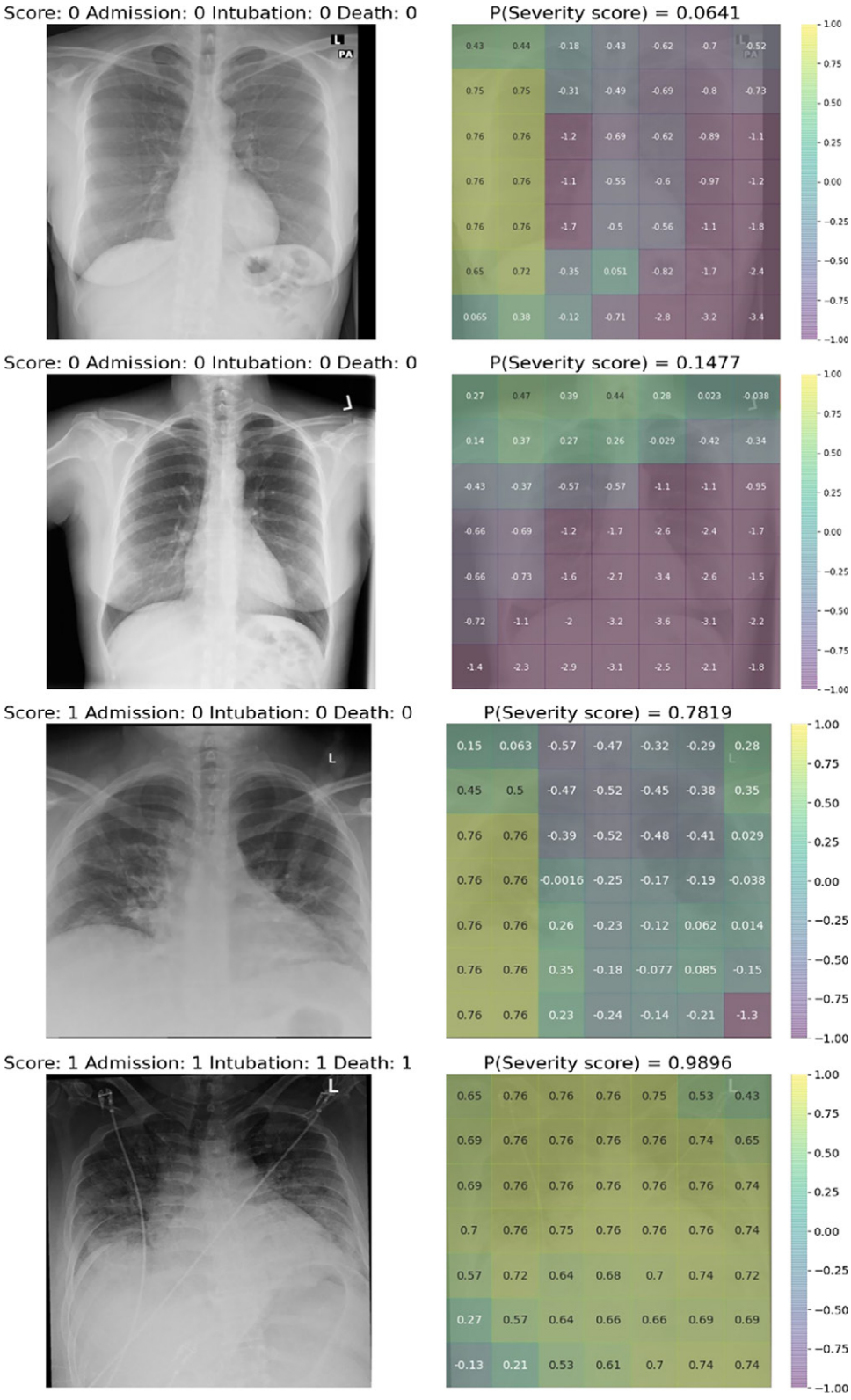



**Figure 6:** Heatmaps generated from the last activation layer of the DenseNet-121 classifier algorithm. As expected, the patient's lower left chest (lower right on the image file), where the heart and gastric bubble are located, does not contribute meaningfully to the prediction output (probability score). Yellow indicates greater contribution to final output, purple less contribution to final output.

containing the key clinical finding due to the convolution architecture of the DL algorithm. Nonetheless, we have planned future studies with additional patients and clinical variables that can help demonstrate both reproducibility and interpretability.

Many AI algorithms that show promising predictive performance do not become integrated into the clinical workflow (40,41). Two authors (Z.A.F., B.S.G.) are part of the COVID Informatics Center in our institution. One of the goals of the center is to deploy these tools in the hospital and integrate them with all available data sources, including electronic health records, such as Epic. Our informatics center works directly with both the clinical staff and the electronic health record staff in our hospital system, and from the beginning, our algorithm was developed with the intention of deployment. As a follow-up study, we are currently investigating the addition of time-course clinical data to the model and are collaborating with potential external contributors. Expansion of the model with longitudinal data and collaboration with external institutions will further generalize the model to a different cohort of admitted patients for whom more clinical laboratory values may have been collected over the course of their hospitalization, a different cohort from this study's patients initially presenting in the ED. The current method presented in this study could form the foundation of incorporating widely available chest radiographs as inputs to more robust prognostication algorithms for determining outcomes in patients with COVID-19.

There are potential challenges to the generalization of this algorithm to the general population. This model did not include patients who did not have real-time reverse-transcription polymerase chain reaction–confirmed COVID-19 in either the training or the test cohort. Thus, this model is inappropriate for predicting COVID-19 when diagnostic test results are not immediately available. This model was trained on only patients aged 21–50 years with test results positive for COVID-19 who were

presumed to have lower occurrences of comorbidities than patients older than 50 years. Nonetheless, our test dataset was diverse, as it contained patients older than 21 years with COVID-19, including older patients with higher risk. Further, our dataset contains data from three hospitals, each of which uses different acquisition devices, and a large proportion of bedside anteroposterior chest radiographs, which are typically inferior to posteroanterior and lateral, but nonetheless are sufficient. We also tested our algorithm on an unseen patient cohort from multiple hospitals that represent diversity of key patient demographics from New York City at a later time point. We recognize that the total number of chest radiographs included in this study ($n$ = 499) is likely too few for general deployment of our algorithm; thus, we are currently acquiring data from similar patient cohorts at external institutions to further validate our algorithm. Nonetheless, the significant increase in performance of our DL model using both chest radiographs and clinical variable data can help inform future prognostication algorithm development.

In summary, we have created a proof-of-concept DL algorithm that was able to predict key clinical outcomes of adult patients with COVID-19, with only the routinely obtained chest radiography and laboratory studies initially acquired in the ED. In doing so, this model validated a chest radiograph severity score that can be used to predict clinical outcomes without any additional clinical variables as inputs. Combining chest radiograph and clinical variables available exclusively from the ED provided the best model performance on predicting intubation and death, better than the models trained on either chest radiography or clinical variables alone. Future work that incorporates additional radiographs and clinical variables acquired at future time points into training the network should further improve the predictive performance. Combination of both imaging and clinical data can help predict clinical outcomes rather than merely the presence of COVID-19 itself and can help triage patients for optimal care.

**Acknowledgments:** The authors express their gratitude to frontline providers and essential workers for their selfless efforts during these unprecedented times.

**Author contributions:** Guarantors of integrity of entire study, Y.J.F.K., D.T., M.F., M.A.C., S.M., A.J., A.B.C.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, Y.J.F.K., D.T., M.A.C., S.Z.M., S.M., C.E., A.J., E.K.O.; clinical studies, Y.J.F.K., D.T., M.A.C., S.Z.M., N.V., C.E., A.J., Y.S.G.; experimental studies, Y.J.F.K., N.V., C.E., A.J., Z.A.F.; statistical analysis, Y.J.F.K., M.F., M.A.C., B.S.G., A.B.C.; and manuscript editing, all authors.

**Disclosures of Conflicts of Interest:** **Y.J.F.K.** disclosed no relevant relationships. **D.T.** disclosed no relevant relationships. **M.F.** disclosed no relevant relationships. **M.A.C.** disclosed no relevant relationships. **S.Z.M.** disclosed no relevant relationships. **S.M.** disclosed no relevant relationships. **N.V.** disclosed no relevant relationships. **C.E.** disclosed no relevant relationships. **A.J.** disclosed no relevant relationships. **A.B.** disclosed no relevant relationships. **Y.S.G.** disclosed no relevant relationships. **M.S.C.** disclosed no relevant relationships. **Z.A.F.** disclosed no relevant relationships. **B.S.G.** disclosed no relevant relationships. **E.K.O.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is a consultant for Google; author's spouse works at Merck. Other relationships: disclosed no relevant relationships. **A.B.C.** disclosed no relevant relationships.

## References

1. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. Radiology 2019;291(1):196–202 [Published correction appears in Radiology 2019;291(1):272.].
2. Li L, Qin L, Xu Z, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 2020;296(2):E65–E71.
3. Huang L, Han R, Ai T, et al. Serial Quantitative Chest CT Assessment of COVID-19: Deep-Learning Approach. Radiol Cardiothorac Imaging 2020;2(2):e200075.
4. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. arXiv [eess.IV] [preprint] http://arxiv.org/abs/2003.05037. Posted March 10, 2020. Accessed April 30, 2020.
5. Choi H, Qi X, Yoon SH, et al. Extension of Coronavirus Disease 2019 on Chest CT and Implications for Chest Radiographic Interpretation. Radiol Cardiothorac Imaging 2020;2(2):e200107.
6. Wong HYF, Lam HYS, Fong AH, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. Radiology 2020;296(2):E72–E78.
7. Kundu S, Elhalawani H, Gichoya JW, Kahn CE. How Might AI and Chest Imaging Help Unravel COVID-19's Mysteries? [Editorial]. Radiol Artif Intell 2020;2(3):e200053.
8. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and Chest Radiography Features Determine Patient Outcomes in Young and Middle-aged Adults with COVID-19. Radiology 2020;297(1):E197–E206.
9. Do S, Song KD, Chung JW. Basics of Deep Learning: A Radiologist's Guide to Understanding Published Radiology Articles on Deep Learning. Korean J Radiol 2020;21(1):33–41.
10. Lu MT, Ivanov A, Mayrhofer T, Hosny A, Aerts HJWL, Hoffmann U. Deep Learning to Assess Long-term Mortality From Chest Radiographs. JAMA Netw Open 2019;2(7):e197416.
11. Cellina M, Panzeri M, Oliva G. Chest Radiography Features Help to Predict a Favorable Outcome in Patients with Coronavirus Disease 2019. Radiology 2020;297(1):E238.
12. Joseph NP, Reid NJ, Som A, et al. Racial and Ethnic Disparities in Disease Severity on Admission Chest Radiographs among Patients Admitted with Confirmed Coronavirus Disease 2019: A Retrospective Cohort Study. Radiology 2020;297(3):E303–E312.
13. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on Chest Radiographs: A Multireader Evaluation of an Artificial Intelligence System. Radiology 2020;296(3):E166–E172.
14. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26(8):1224–1228.
15. Wu Q, Wang S, Li L, et al. Radiomics Analysis of Computed Tomography helps predict poor prognostic outcome in COVID-19. Theranostics 2020;10(16):7231–7244.
16. Li MD, Arun NT, Gidwani M, et al. Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Convolutional Siamese Neural Networks. Radiol Artif Intell 2020;2(4):e200079.
17. Abdulaal A, Patel A, Charani E, Denny S, Mughal N, Moore L. Prognostic Modeling of COVID-19 Using Artificial Intelligence in the United Kingdom: Model Development and Validation. J Med Internet Res 2020;22(8):e20259.
18. Subudhi S, Verma A, Patel AB. Prognostic machine learning models for COVID-19 to facilitate decision making. Int J Clin Pract 2020;74(12):e13685.
19. Liu F, Zhang Q, Huang C, et al. CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients. Theranostics 2020;10(12):5613–5622.
20. Cardinale L, Priola AM, Moretti F, Volpicelli G. Effectiveness of chest radiography, lung ultrasound and thoracic computed tomography in the diagnosis of congestive heart failure. World J Radiol 2014;6(6):230–237.
21. Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. MMWR Morb Mortal Wkly Rep 2020;69(15):458–464.
22. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv [cs.CV.] [preprint] http://arxiv.org/abs/1711.05225. Posted November 14, 2017. Accessed April 30, 2020.
23. Deng J, Dong W, Socher R, Li L, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, June 20–25, 2009. Piscataway, NJ: IEEE, 2009; 248–255.

24. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv [cs.CV] [preprint] http://arxiv.org/abs/1608.06993. Posted August 25, 2016. Accessed April 30, 2020.
25. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG] [preprint] http://arxiv.org/abs/1412.6980. Posted December 22, 2014. Accessed April 30, 2020.
26. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 2921–2929.
27. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. arXiv [cs.CV] [preprint] http://arxiv.org/abs/1807.00431. Posted July 2, 2018. Accessed April 30, 2020.
28. DiCiccio TJ, Efron B. Bootstrap Confidence Intervals. Stat Sci 1996;11(3):189–228.
29. Hall P, Hyndman RJ, Fan Y. Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika 2004;91(3):743–750.
30. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, eds. Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, vol 8190. Berlin, Germany: Springer, 2013; 451–466.
31. Fine MJ, Hough LJ, Medsger AR, et al. The hospital admission decision for patients with community-acquired pneumonia. Results from the pneumonia Patient Outcomes Research Team cohort study. Arch Intern Med 1997;157(1):36–44.
32. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med 1997;336(4):243–250.
33. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020;121:103792.
34. Li T, Han Z, Wei B, Zheng Y, Hong Y, Cong J. Robust Screening of COVID-19 from Chest X-ray via Discriminative Cost-Sensitive Learning. arXiv [eess.IV] [preprint] http://arxiv.org/abs/2004.12592. Posted April 27, 2020. Accessed April 30, 2020.
35. Yoon SH, Lee KH, Kim JY, et al. Chest Radiographic and CT Findings of the 2019 Novel Coronavirus Disease (COVID-19): Analysis of Nine Patients Treated in Korea. Korean J Radiol 2020;21(4):494–500.
36. Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. J Thorac Imaging 2020;35(3):W87–W89.
37. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–510.
38. Hwang EJ, Park CM. Clinical Implementation of Deep Learning in Thoracic Radiology: Potential Applications and Challenges. Korean J Radiol 2020;21(5):511–525.
39. Lee JG, Jun S, Cho YW, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol 2017;18(4):570–584.
40. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019;28(3):238–241.
41. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):195.