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

Article Type: Original Research

Summary Statement

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 prognosticated 30-day intubation and death trained on only routine chest radiograph (intubation area under the receiver operating curve [AUC] 0.66, death AUC 0.59) or clinical laboratory values (intubation AUC 0.64, death AUC 0.59) better than a naïve classifier.

■ Performance of prediction of intubation (AUC, 0.88) and death (AUC, 0.82) increased when the model was trained with initial chest radiograph *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 to 50, generalized to a pseudo-prospective test set that also contained older patients aged greater than 50.

Abbreviations

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

Abstract

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

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

Results: The model trained on the CXR severity score produced the following areas under the receiver operating characteristic (AUCs): 0.80 (0.73,0.88) for the CXR severity score, 0.76 (0.68,0.84) for admission, 0.66 (0.56,0.75) for intubation, and 0.59 (0.49,0.69) for death. The model trained on clinical variables produced the following AUCs 0.64 (0.55,0.73) for intubation and 0.59 (0.50,0.68) for death. Combining CXR and clinical variables increased AUC of intubation and death to 0.86 (0.79,0.96) and 0.82 (0.72,0.91), respectively.

Conclusion: Combination of imaging and clinical information improves outcome predictions.

Introduction

Artificial intelligence has demonstrated promise in facilitating triage in radiology departments due 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 its clinical value in screening for or diagnosing coronavirus disease 2019 (COVID-19) predominantly when employed using chest CT.[2–4] In clinical practice, however, chest radiography (CXR) is the primary and often only imaging modality that is obtained for patients with COVID-19, particularly in health systems with limited resources.[5] Although the value of CXR in diagnosing COVID-19 might be limited by its reported low sensitivity, it may be useful in the prognostication of COVID-19 positive patients.[6–8]

Deep learning (DL) is a type of artificial intelligence in which data is processed iteratively through multi-layered neural networks to automatically extract high level features from raw data input. This recursive method allows for programs to discern patterns without explicit human guidance.[9] Recently, a DL algorithm has been reported to accurately predict long-term outcomes from single CXRs from patients with prostate, lung, colorectal, and ovarian cancer.[10] Another cohort from Italy showed data that supported the role of CXR 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 need for assisted ventilation.[11] There is a growing body of literature in using CXR that showed increased severity was associated with worse outcomes for all patients.[12] Some DL algorithms incorporating CT and CXR data have been used to aid in screening and diagnosis of COVID-19, and one study used CT to predict poor prognostic outcomes in patients with COVID-19.[2,13–15] Recently, a model that predicts a “Pulmonary X-ray Severity score” based on CXRs from patients with COVID-19 was published.[16] Nonetheless, the potential of DL algorithms, especially those that have been trained with CXRs 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 CXR image as input has been largely unexplored.[8] In fact, many currently published prognostication algorithms use only clinical variables and do not use imaging data as input or use only CT, a modality less widely available and less frequently obtained than CXR.[17–19]

The presence of comorbidities such as lung and heart disease can potentially confound CXR interpretation of patients with COVID-19 pneumonia, which may decrease the predictive ability of DL.[20] In this context,

therefore, the generation of predictive CXR interpretations may be more valid in patients under 50 years of age, who have a lower prevalence of such conditions. While COVID-19 affects all ages, the younger population still comprises a considerable proportion of affected patients.[21] Thus, testing for generalizability of prognostication algorithms for patients with COVID-19 is important for deployment of DL to appropriate patient populations.

In this study, we propose a proof of concept model aimed to demonstrate that a DL algorithm can take only the initial CXR, an imaging study that the 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 CXR *or* clinical variables alone to that of the model trained on *both* CXR and clinical variables evaluate individual contribution of CXR or clinical variables to the prognostication and to test for 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 pathologies, including pneumonia, on CXRs.[22] We hypothesized that training the convolutional neural network with image input and the associated CXR severity score, which was previously reported and validated in Toussie et al, is as effective as training with the image input and the associated clinical outcome of admission as labels.[8] We then tested this model to generate a model severity score that is 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 aged greater than 50, who presented at different time points to predict admission, intubation, and mortality. We also supplemented the model with standard lab tests available at the initial ED encounter to increase the model performance.

Materials and Methods

Patient Selection

To collect the patient cohort for this institutional review board approved retrospective cohort study with waived written consent, we used the MONTAGE™ Search and Analytics Platform and extracted radiology information system data from all CXR examinations performed in the ED setting from March 10 - 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 obtained HIPAA approval. Using the obtained cohort, we then extracted relevant clinical and laboratory data from the electronic medical record

(EMR). The resulting radiology information system dataset contained 4738 ED encounters. The exclusion criteria included greater than 50 or less than 21 years of age (*n* = 3163), duplicate CXR of the same patient (*n* = 81), patients with unconfirmed COVID-19 real-time reverse transcription polymerase chain reaction positivity (*n* = 1101), presentations unrelated to COVID-19 (*n* = 2), unevaluable CXR (*n* = 1), and inaccessible clinical data (*n* = 1). All 338 patients from the original Toussie et al. clinical study from *Radiology* were included in the train and the validation dataset.[8] 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 CXR directly.

We randomly assigned the included CXRs from March 10 - 26 (*n* = 338) to either the training set (*n* = 283; 84%) or the validation set (*n* = 55; 16%) for the DL model. In the training set, 73.5% (208 of 283) of the radiographs were acquired portably with anteroposterior views and 26.5% (75 of 283) were acquired with posteroanterior and lateral views. In the validation set, 76.4% (42 of 55) radiographs were acquired with portable anteroposterior views and 23.6% (13 of 55) with posteroanterior and lateral views. We used only frontal radiographs for model training. The included CXRs from March 27 - 29, 2020 (*n* = 51) were assigned to compile a held out test set from a different time period (**Figure 1**). There were a total of 161 patients included within the test set. A total of 51 of these patients were between the ages of 21 and 50 years, while 110 patients were aged greater than 50 years. These 110 patients were added to test for the generalizability of the model in older patients at a greater risk. Within the test set of patients aged 21-50 years (*n* =51), 68.6% (35 of 51) radiographs were acquired with portable anteroposterior views and 31.4% (16 of 51) posteroanterior and lateral views. Within the test set of patients aged greater than 50 years (*n* = 110), 96 (87.3%) radiographs were acquired with portable anteroposterior views and 14 (12.7%) with posteroanterior and lateral views (**Table 2)**. We used only frontal radiographs for model inference.

Data Collection

Two fellowship-trained cardiothoracic radiologists, blinded to patient history other than COVID-19 positivity, independently examined the initial CXR for opacities to generate a total severity score (CE with 26 years of experience and SG with 1 year of experience). Each lung was divided into three zones - upper, middle, and lower zones - and a binary score of 0 (no opacity) or 1 (opacity) was assigned to each lung zone (**Figure E1**

[supplement]).[8] For model training, only the lung zones that both radiologists agreed to contain opacity was given the final opacity label (score of 1); otherwise, the lung zones were noted as normal (score of 0). CXRs with scores of 2 or above (out of 6) were categorized as severe for purposes of the training algorithm. Any admission, intubation, or death in the 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 1024 x 1024 resolution. 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 mean and standard deviation. They were stored as HDF5 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 pre-trained 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 (**Figure 2**).[25] We empirically tested for the best learning rate from 1×10^{-2} to 1×10^{-10} in logarithmic increments (1×10^{-2}, 1×10^{-3}, … , 1×10^{-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: C-reactive protein, white blood cell count, D-dimer, lactate, lactate dehydrogenase, creatinine, eGFR, troponin, aspartate aminotransferase, glucose, systolic and diastolic blood pressures. We used mean imputation for any unavailable lab values. For model training with clinical variables alone, we used fully connected layers. For model training with both CXR and 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 CXR and its severity score from above (**Figure 2**).

Model Evaluation

We selected the model from the training with either the CXR 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 CXR image as input was used to calculate the area under the receiver operator characteristic curve (AUCs) for four different classes: CXR severity scores, admissions, intubations, and deaths. For example, a DL algorithm generated score above .65 predicts admission, above .80 predicts intubation, and above .90 predicts death. To account for variable prevalence among classes, we designated classes that had a prevalence greater than or equal to 40% in our cohort as “majority class” while those with a prevalence lower than 40% were designated as “minority class”. Severe CXRs and admissions were thereby majority classes while intubation and death were minority classes. We then plotted the precision-recall (PR) curve to evaluate the model performance for minority classes which were not used as part of the training. We used the discriminative localization methods previously described to generate heatmaps that describe which parts of the radiographs were contributing the most to the prediction algorithm.[26,27] The source code used in this paper 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 patient race, patient sex, smoking history, hospital site, and comorbidities was performed using chi-squared 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 for accuracy and F1 score calculations. 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 are 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 CXR, 30-day admission, intubation, and death).

Results

Patient Demographics

Overall, 499 patients and their CXRs were used between the training, validation, and test sets with a diverse patient population. Of all 499 CXRs that were scored, 41 CXRs (8.2%) had severity scores 2 or above given by one of the two reviewers but not when the severity score was calculated with concordant scores. The remaining 458 CXRs (91.8%) had been correctly categorized as severe (scores 2 or above) or not severe (scores 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 CXRs, 271 (54.3%) were admitted, 73 (14.8%) were intubated, and 51 (10.2%) expired. Additionally, there were 53 (10.6%) with asthma, 3 (0.6%) with COPD, 105 (21.0%) with hypertension, 73 (14.6%) with diabetes mellitus, 7 (1.4%) with HIV, 18 (3.6%) with cancer, 23 (4.6%) with chronic kidney disease, 25 (5.0%) with coronary artery disease, and 5 (1%) with atrial fibrillation. The datasets differed significantly with regards to age (due to inclusion of patients aged greater than 50 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, there time from initial CXR to intubation had an average of 3.7 days and a median of 3 days (range: 0 to 12 days).

The training, validation, and test datasets consisted of 283, 55, and 161 patients, respectively. Severe CXR, admission, intubation, and death data for these datasets are found in **Table 2.** The subset of the test set of 51 patients aged 21 to 50 years had 34 (66.7%) severe CXRs, 34 (66.7%) admissions, 10 (19.6%) intubations, and seven (13.7%) deaths (**Table 2**). Of the 499 CXRs that were scored, 41 CXRs (8.2%) had severity scores 2 or above given by one of the two reviewers but not when the severity score was calculated with concordant scores. The remaining 458 CXRs (91.8%) had been correctly categorized as severe (scores 2 or above) or not severe (scores 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 1×10^{-5}, the b1 decay 0.99, b2 decay 0.9999, and weight decay 1×10^{-5} after 10 epochs of training. Both training with the CXR severity scores or the admission status converged to the best model as evaluated by the validation loss (**Figure E2 [supplement]**). Initially, iterations of training

demonstrated low AUC for predicting death in the validation set, but increased with additional iterations (**Figure E2 [supplement]**).

Prediction of Independent Clinical Outcome Variables

After selection of the best model based on the minimum validation loss (**Figure 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 CXR severity scores or the model trained with admissions, was then used to generate AUC values for the CXR 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 CXR severity score produced the following AUCs: 0.80 (95% CI: 0.73, 0.88) for CXR severity score, 0.76 (0.68, 0.84) for admission, 0.66 (0.56, 0.75) for intubation, and 0.59 (0.49, 0.69) for death (**Figure 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 the admission status produced the following AUCs: 0.70 for CXR 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 (**Figure 3**).

The precision-recall (positive predictive value - sensitivity) curve suggests that the performance was better on majority classes (CXR severity score and admission status) than on minority classes (intubation status and death). The accuracy on 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 (30% and 26% for 30-day intubation and death, respectively) (**Table 3**, **Figure 4**). Further, the performance of negative predictive value and specificity (ie precision for the minority class) was better at predicting lack of intubation or survival than a naive classifier (**Figure E3 [supplement]**).

The model performance on the prediction on intubation and death increased when trained with clinical variables from electronic health records and with intubation status as the target label (**Figure 5**). 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 aged greater than 21 years. The combined model performed better than the model trained on clinical variables alone as well. As expected, the model performed better for the young adults aged 21-50 years, but still demonstrated clinically useful results for the older

patients aged greater than 50 years in the test set. At our selected operating point for intubation that prioritizes recall or sensitivity greater than 80%, F1-score still remains high above 65%.

The heatmap results indicate that the inferior left of the patient's chest anatomy (right side of the radiograph) that contains the heart and the gastric bubble contribute less to the radiograph compared to the rest of the radiograph. The absolute value of the model output (given as a probability) increases with worse clinical outcomes (**Figure 6**).

Discussion

We hypothesized that a DL model could predict prognosis of adult patients with COVID-19 based solely on routinely available imaging (CXR) 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 aged greater than 50 in the test set to assess for the generalizability of the model in older, higher-risk patients. Using a previously successful DL classification algorithm, DenseNet-121, we trained the model successfully with the CXR and the associated severity score or the 30 day admission status. This trained model could then take unseen CXR from another time period to predict the 30 day admission status, intubation status, and survival, despite the differences of patient age and outcomes in the test set compared to that of the training and validation sets. We also trained a model with clinical variables alone and compared the models trained on either CXR or clinical variables only to a model trained with both CXR and clinical variables. The combined model had the best performance.

Fine et al 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.[31] Furthermore, CURB-65 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 demonstrated that the severity of opacities on chest radiographs does predict outcomes in COVID-19 pneumonia.[8] The severity of opacity on the presentation chest radiograph is an important objective assessment of the severity of disease that can be used to guide physician decisions on whether a patient needs to be admitted or can safely be discharged and managed at home. We used CXRs 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 COVID-19 positive patients. While COVID-19 is still rapidly spreading across the United States and overwhelming hospitals, a quick tool that can provide accurate prognostication for COVID-19 can help appropriately allocate resources (eg inpatient hospital beds, ventilators) for subsequent management is vital.

The advantage of the design within this study lies in the ability for a DL model to predict clinical outcomes rather than screening for or confirming 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 CXRs used in the model and the different time frame of the test cohort (ie a pseudo-prospective trial) suggest a higher likelihood of generalizability. While surveys of ED physicians do not typically report CXR 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 COVID-19 patients and may serve as a first-pass triaging process to alert radiologists and clinicians of higher-risk patients who will likely require hospitalization.[31] This prioritization of care can be readily adopted within existing clinical workflows and lead to validation in actual clinical practice, thereby addressing the common challenges and criticisms of existing artificial intelligence research in medicine.[37,38]

This study confirms that the CXR severity score can be used to train a network that predicts clinical outcomes, including need for hospitalization, intubation, and mortality. There are multiple benefits of using the initial CXR severity score rather than the outcome of interest in the prediction model. Firstly, the model post deployment requires only the initial CXR to provide prognostic predictions without 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 CXRs from the ED, there is no need for a 30-day follow-up. Additionally, the CXR scores can be incorporated into the model without having to wait 30 days to confirm the absence of an admission event. It is possible that the patients are admitted for other non-respiratory reasons, whereas the severity score that indicates opacity in CXR lung zones more directly correlates to potential intubation. Therefore, the dataset can be expanded immediately with the widespread availability of the CXR and the initial lab 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 it is trained with either the severity scores or the clinical outcomes.

The precision and recall of the predictions, based on CXR alone, for intubation and death did not have as high of performance as those for CXR severity scores and admission because of the relative scarcity of these events. Nonetheless, this model still predicted 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 Toussie et al. study that the findings from the initial CXR obtained in the ED contains information that will enable physicians to better predict need for hospitalization and help ensure the appropriate patients are admitted versus discharged.[8] The model trained only on CXR had similar AUCs for intubation and death prediction as those of the model trained only with clinical variables first obtained from the ED. The model trained only with clinical variables had a low true positive rate and high false positive rate at high cutoffs (left side of receiver operator characteristic curve, **Figure 5**). That is, with 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 a single model. Our model, which uses only the information from the initial ED encounter from standard imaging and routinely ordered lab tests, may be used to help guide hospitalization decisions of patients with COVID-19 and inform ED physicians on the risks of their patients developing poor outcomes later in the disease course. The timeline to prognostication is clinically relevant, given that our patient cohort that required intubation had a median of 3 days from the first CXR to intubation.

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 parts of the radiograph were not contributing significantly to the final output. Of note, heatmaps do not indicate which anatomical regions and their qualities are truly contributing to the prediction. Further, the regions generated by the heatmaps may be of different sizes than the actual subregion 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] The authors 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 informatic center works directly with both the clinical and the electronic health record staff in our hospital system, and our algorithm was developed with the intention of deployment from the beginning. 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 institutes will further generalize the model to a different cohort of admitted patients who may have more clinical lab values collected over the course of their hospitalization, a different cohort from this study's initially presenting patients in the ED. The current method presented in this study could form the foundation of incorporating widely available CXRs 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 that 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 prediction of COVID-19, when diagnostic testing is not immediately available. This model was trained on only COVID-19 positive patients aged 21 and 50 years who were presumed to have lower occurrences of comorbidities. Nonetheless, our test dataset was diverse as it contained patients with COVID-19 of all ages greater than 21 years that included older, higher at-risk patients. Further, our dataset contains data from three hospitals that each use different acquisition devices and a large proportion of portable anteroposterior CXRs, a technique that is typically inferior to posteroanterior and lateral views, but nonetheless sufficient. We also tested our algorithm on an unseen patient cohort at a later time point from multiple hospitals that represent diversity of key patient demographics from New York City. We recognize that 499 total CXRs included in this study is likely too few for general deployment of our algorithm, and 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 CXR 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 CXR and laboratory studies initially acquired in the ED. In doing so, this model validated a CXR severity score that can be used to predict clinical outcomes without any additional clinical variables as inputs. Combining CXR and clinical variables available exclusively from the ED had the best model performance on predicting intubation and death, better than the models trained on either CXR 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 the presence of COVID-19 itself, and can help triage patients for the best patient care.

Acknowledgments

The authors would like to acknowledge funding support from the RSNA Medical Student Research Grant and the T32 NIH T32 Medical Scientist Training Program Grant. Most importantly, we express our gratitude to front-line providers and essential workers for their selfless efforts during these unprecedented times.

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.

2. Li L, Qin L, Xu Z, et al. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. *Radiology*. Published online March 19, 2020:200905.

3. Huang L, Han R, Ai T, et al. Serial Quantitative Chest CT Assessment of COVID-19: Deep-Learning Approach. *Radiology: Cardiothoracic 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 [eessIV]*. Published online March 10, 2020. http://arxiv.org/abs/2003.05037

5. Choi H, Qi X, Yoon SH, et al. Extension of Coronavirus Disease 2019 (COVID-19) on Chest CT and Implications for Chest Radiograph Interpretation. *Radiology: Cardiothoracic Imaging*. 2020;2(2):e200107.

6. Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients. *Radiology*. Published online March 27, 2019:201160.

7. Kundu S, Elhalawani H, Gichoya JW, Kahn CE. How Might AI and Chest Imaging Help Unravel COVID-19's Mysteries? *Radiology: Artificial Intelligence*. 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 Age Adults with COVID-19. *Radiology*. Published online May 14, 2020:201754.

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 Radiograph Features Predict a Favorable Outcome in Patients with COVID-19. *Radiology*. Published online June 2, 2020:202326.

12. Joseph NP, Reid NJ, Som A, et al. Racial/Ethnic Disparities in Disease Severity on Admission Chest Radiographs among Patients Admitted with Confirmed COVID-19: A Retrospective Cohort Study. *Radiology*. Published online July 16, 2020:202602.

13. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on the Chest Radiograph: A Multi-Reader Evaluation of an AI System. *Radiology*. Published online May 8, 2020:201874.

14. Mei X, Lee H-C, Diao K-Y, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med*. Published online May 19, 2020. doi:10.1038/s41591-020-0931-3

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. *Radiology: Artificial Intelligence*. 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, B Patel A. Prognostic machine learning models for COVID-19 to facilitate decision making. *Int J Clin Pract*. Published online August 18, 2020: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 [csCV]*. Published online November 14, 2017. http://arxiv.org/abs/1711.05225

23. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009:248-255.

24. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. *arXiv [csCV]*. Published online August 25, 2016. http://arxiv.org/abs/1608.06993

25. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv [csLG]*. Published online December 22, 2014. http://arxiv.org/abs/1412.6980

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*. ; 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 [csCV]*. Published online July 2, 2018. http://arxiv.org/abs/1807.00431

28. DiCiccio TJ, Efron B. Bootstrap Confidence Intervals. *Stat Sci*. 1996;11(3):189-212.

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: *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg; 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*. doi:10.1016/j.compbiomed.2020.103792

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 [eessIV]*. Published online April 27, 2020. http://arxiv.org/abs/2004.12592

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 J-G, Jun S, Cho Y-W, et al. Deep Learning in Medical Imaging: General Overview. *Korean J Radiol*. 2017;18(4):570-584.

40. Yu K-H, 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.

Tables

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

	Overall	Training	Validation	Test	*P* value
Total	499	283	55	161	NA
Men(%)	308 (62)	174 (62)	36 (65)	98 (61)	.83
Age, years	42 (34, 50)	38 (31, 45)	41 (35, 44)	60 (46, 70)	< .001
Race (%)					
White	99 (20)	59 (21)	12 (22)	28 (17)	.1
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 (%)					
Normal	135 (29)	81 (28)	11 (20)	43 (33)	.01
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 (%)					
No	327 (66)	186 (66)	37 (67)	104 (65)	.15
Former	58 (12)	25 (9)	5 (9)	28 (17)	
Other or unknown	86 (17)	54 (19)	10 (18)	22 (13)	
Yes	28 (6)	18 (6)	3 (6)	7 (4)	
Site (%)					
Manhattan	201 (40)	118 (42)	25 (46)	58 (36)	.39
Brooklyn	154 (31)	90 (32)	12 (22)	52 (32)	
Queens	144 (29)	75 (27)	18 (33)	51 (32)	

Note.— Continuous variables shown as mean (interquartile range) and categorical variables are shown as number of patients (percentage). BMI = body mass index

Table 2. Distribution of Imaging Modality, Severe CXR, and Clinical Outcomes.

	Total (*n* = 499)	Training (*n* = 283)	Validation (*n* = 55)	Test (*n* = 161)
Modality				
Portable	381 (76.4%)	208 (73.5%)	42 (76.4%)	131 (81.4%)
PA and Lat	118 (23.6%)	75 (26.5%)	13 (23.6%)	30 (18.6%)
Severe CXR 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.— Only frontal views from posteroanterior (PA) and lateral acquisitions were used for training. The test set includes 110 patients that were aged greater than 50 years of age. CXR = chest radiograph

* The 30-day intubation value in the test set excludes five patients aged greater than 50 who were indicated as “Do Not Intubate”.

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 to 50 or aged greater than 50. All values are percentages (95% CI).

	All patients (*n* = 161)			Patients aged 21-50 (*n* = 51)			Patients aged > 50 (*n* = 110)		
	Naive Classifier	Trained on Scores	Trained on Admissions	Naive Classifier	Trained on Scores	Trained on Admissions	Naive Classifier	Trained on Scores	Trained on Admissions
A. 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)
B. 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)
C. 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)
D. 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)

Figure Legends

Figure 1. Patient inclusion and exclusion criteria. COVID-19 = coronavirus disease 2019, CXR = chest radiograph, ED = emergency department, MRN = medical records number.

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

Figure 3. Receiver operating characteristic (ROC) curves of the test set based on two different training schemes. The areas under the ROC curves (AUCs) do not differ between training on severity score or admission status. The 95% confidence intervals for the AUCs are indicated by the brackets, shown as [lower bound, upper bound]. Operating point was selected for high sensitivity (recall), which was then used for accuracy and F1 score calculations.

Figure 4. Precision versus recall curves for four prediction categories. Chest radiograph severity score and admission status were used for training and well balanced. Intubation status and mortalities were minority classes and 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.

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 and their severity score. The AUC for predicting intubation increased from 0.66 to 0.88 and for predicting death increased from 0.59 to 0.82. At our selected operating point, the sensitivity remained high while achieving a good F1-score. Intervals indicate 95% CIs. Five patients who were “do not intubate” in the test set were excluded from intubation data analysis.

Figure 6. Heatmaps generated from the last activation layer of the DenseNet-121 classifier algorithm. As expected, the patient's lower left (lower right on the image file) where the heart and the gastric bubble is located does not contribute significantly to the prediction output (probability score).