A DEEP LEARNING MODEL TO PREDICT THE NEED FOR MECHANICAL VENTILATION USING CHEST X-RAY IMAGES IN HOSPITALIZED COVID-19 PATIENTS

Anoop Kulkarni [1,2,*], Ambarish M. Athavale [3,*], Ashima Sahni [4], Shashvat Sukhal [5], Abhimanyu Sahni [6], Mathew Itteera [3], Sara Zhukovsky [7], Jane Vernik [3], Mohan Abraham [3], Amit Joshi [3], Amatur Amarah [3], Juan Ruiz [3], Peter D. Hart [3], Hemant Kulkarni [2,8]

[1] Innotomy Consulting, Bengaluru, India

[2] Lata Medical Research Foundation, Nagpur, India

[3] Division of Nephrology, Department of Medicine, Cook County Health, Chicago, Illinois, USA

[4] Division of Pulmonary, Critical Care, Sleep, and Allergy, University of Illinois Hospital and Health Sciences System, Chicago, Illinois, USA

[5] Division of Pulmonary and Critical Care, Department of Medicine, Cook County Health, Chicago, Illinois, USA

[6] Division of Cardiology, Department of Medicine, Cook County Health, Chicago, Illinois, USA

[7] Rush University Medical Center, Chicago, IL, USA

[8] M&H Research, LLC, San Antonio, Texas, USA

*, these authors contributed equally

Running Title: Deep learning from X-rays of COVID-19 patients

Word Count: 3,272; **Number of Tables: 1**; **Number of Figures: 4**

Corresponding author:
Hemant Kulkarni, MD
12023 Waterway Rdg,
San Antonio, TX 78249, USA
Phone: +1 (210) 602 5537
Email: hemant.kulkarni@mnhresearch.com

DISCLOSURES: None of the authors have a conflict of interest to disclose.

ABSTRACT

Purpose: Early identification of a potentially deteriorating clinical course in hospitalized COVID-19 patients is critical since there exists a resource-demand gap for the ventilators.

Materials: We aimed to develop and validate a deep learning-based approach to predict the need for mechanical ventilation as early as at the time of initial radiographic evaluation. We exploited the well-established DenseNet121 deep learning architecture for this purpose on 663 X-ray images derived from 528 hospitalized COVID-19 patients. Two Pulmonary and Critical Care experts blindly and independently evaluated the same X-ray images for purpose of validation.

Results: We found that our deep learning model predicted the need for ventilation with a high accuracy, sensitivity and specificity (90.06%, 86.34% and 84.38%, respectively). This prediction was done approximately three days ahead of the actual intubation event. Our model also outperformed two Pulmonary and Critical Care experts who evaluated the same X-ray images and provided an incremental accuracy of 7.24–13.25%.

Conclusion: Our deep learning model accurately predicted the need for mechanical ventilation early during hospitalization of COVID-19 patients. Until effective preventive or treatment measures become widely available for COVID-19 patients, prognostic stratification as provided by our model is likely to be highly valuable.

KEYWORDS: deep learning, chest radiograph, COVID-19, mechanical ventilation

INTRODUCTION

Coronavirus Disease-19 (COVID-19) is a global pandemic which has caused an estimated 20 million infections and 738,668 deaths within a span of just over seven months. [1] Strikingly, the cumulative COVID -19 hospitalization rate is 137.6 per 100,000 infections. [2] COVID-19 can affect major organ systems such as lungs, heart, kidney and brain but the morbidity and mortality in COVID-19 patients is primarily due to lung infection from a pneumonic process. Thus, a significant number of COVID-19 patients need supportive care such as intravenous fluid administration, supplemental oxygen. Further, as many as 32% of hospitalized COVID-19 patients need admission to an intensive care unit [3] and mechanical ventilation [4, 5]. This has caused a great strain in hospital resources in certain geographic regions with high rates of COVID-19 infection. For example, at the height of COVID-19 pandemic in Wuhan and New-York, there were concerns of the health system being overwhelmed from the sheer number of patients requiring hospitalization. In Wuhan, a temporary COVID-19 facility was built and in New-York, a United States Navy hospital ship was dispatched to help cope with the number of patients requiring hospitalization[6, 7]. Since then, COVID-19 has now spread and threatens to overwhelm hospital systems in Texas, Florida and Arizona with the possibility of a second wave in the fall and winter coinciding with the annual rise of influenza. In the absence of a vaccine or effective anti-viral treatment, hospital systems will need to be prepared for an increase in hospitalization rates, ICU admissions and need for mechanical ventilation.

It has been observed that many COVID-19 patients experience a worsening of shortness of breath and need for supplemental oxygen or mechanical ventilation during the second week of

the illness[8]. However, not every patient who is hospitalized with COVID-19 infection needs mechanical ventilation or ICU level of care. Thus, a tool that can effectively predict the need for mechanical ventilation would ensure a better triage at initial point of contact with healthcare system and enable better allocation of healthcare resources by avoiding unnecessary hospitalizations. This was the motivation for the present study.

Deep learning is a branch of artificial intelligence that has shown great promise in diagnosis and prognosis of various health conditions. Specifically, in the context of COVID-19 disease, a deep learning analysis of chest radiograph was able to identify patients with COVID-19 infection with more than 90% accuracy. [9] Also, Wang et al [10] were able to stratify patients in high risk and low risk groups by a deep learning analysis of lung computed tomography (CT) images. In this study, we focused on using the information contained within chest X-ray images to predict the need for mechanical ventilation. A chest radiograph has practical advantages over CT scans in being more readily available especially in resource-challenged scenarios and less risk of equipment contamination. Indeed, a chest radiograph was performed for every patient with COVID-19 evaluated in our hospital emergency room. Here, we present a deep learning analysis of Chest radiograph of patients with COVID-19 to predict need for mechanical ventilation.

MATERIAL AND METHODS

Study participants

The clinical and image data for this study were collected at the John H. Stroger, Jr Hospital of Cook County, Chicago, IL. All COVID-19 patients who were admitted to the study center

between March 15, 2020 and May 31, 2020 and followed up till the censoring date of June 16, 2020 were included. On the last day, 7 (1.3%) of the patients were still in hospital all of whom had completed at least 16 days of inpatient follow-up. The study cohort was identified in two: first, all confirmed COVID-19 cases were identified and second, only the new inpatients were selected from those identified in step one. COVID-19 positivity was confirmed for all patients using the polymerase chain reaction for the RdRp and N genes of the SARS-CoV-2 virus. Clinical data of these patients was collected by chart reviews. The study did not require informed consent from the patients as the data was retrospectively collected after de-identification. The study was approved by the Institutional Review Board of the Cook County Health, Chicago, IL.

Chest X-ray images

All patients with symptoms suggestive of a possible infection with SARS-CoV-2 underwent portable antero-posterior chest X-ray assessment at the study center. The protocol followed was as follows: Ensuring appropriate isolation and distancing practices the X-ray images were acquired in upright or near-upright posture. Images were saved in dicom and jpg format and were manually scrubbed to remove all identifiable information.

Data preprocessing

The acquired X-ray images were first resized to 224 x 224 pixels and then center cropped as required for many deep learning networks that use convolutional layers to parse out image features. To ensure robustness in training and validation of the deep learning network, we undertook two steps in data preprocessing. First, we augmented each image using a random

combination of right- or left-rotation (max 30°), random cropping and random lighting. These augmentations permitted us to use different variations of the original image for training the deep learning algorithm thereby reducing the potential overfitting. Second, since the patients who needed mechanical ventilation in the study dataset represented a minority class, for training the network we first oversampled the number of ventilated patients so as to achieve a class balance of ~50% of X-ray images for ventilated and non-ventilated patients in the training sample. Combination of the first and second steps in data preprocessing yielded a set of 1320 X-ray images from the ventilated patients and 1200 X-ray images from non-ventilated patients. This set of 2520 images was used for network training.

Network architecture

The established and validated CheXNeXt deep learning algorithm [11] as well as the PXR network [12] are based on the DenseNet121 [13] architecture. While the CheXNeXt predicts one or more of 14 lung pathologies from an X-ray image of the chest, the PXR network scores an X-ray image for severity of acute respiratory distress syndrome (ARDS). We used the same backbone for our proposed prognosticator algorithm. The architecture of a DenseNet121 network is shown in Figure 2B. Briefly, the DenseNet121 represents a series of convolutional operations on the image array (size 224 x 224 pixels) and is characterized by a serial combination of 4 dense blocks (D1-D4, Figure 2B) interspersed with 3 transitional blocks (T1-T3, Figure 2B). Each dense block is, in turn, a serial combination of densely connected convolutional layers such that each succeeding layer receives inputs from all preceding layers. The total number of hidden layers in a DenseNet121 network are 121 (hence the name) and the output is

typically given as a multi-probability array which is subjected to a softmax function to obtain likely classifications. In our case, since the outcome (need for mechanical ventilation) was binary, we changed the last layer to a sigmoid function (equivalent to a logit function in logistic regression) as shown in Figure 2B. We used this modified DenseNet121 network architecture in our study.

Network Training and Validation

We used the Tensorflow 2.2.0 (https://www.tensorflow.org/) and Keras 2.3.0-tf framework (https://keras.io/) for model training and evaluation. The Jupyter notebook congaing all the Python code is available with the authors and will be shared upon receipt of reasonable request. Training of the model was done on all the layers of DenseNet121 (i.e. no layers were frozen) with the following pre-specifications: batch size: 32, optimizer: stochastic gradient descent (SGD), loss function: binary cross-entropy, learning rate: 0.003, epoch
s per cycle length: 4 (with plateaued loss) and cycle length: 10. Model that provided the best validation accuracy was selected as the final model.

X-ray evaluation by Pulmonary and Critical Care (PCC) Experts

Two experienced experts from the field of Pulmonary and Critical Care (PCC) evaluated all the X-ray images included in the independent test set (153 images on 118 patients). Both the PCC experts answered the following question for each X-ray image evaluated: “Based on this X-ray image, do you think that this COVID-19 patient will need to be mechanically ventilated during the index hospitalization?” These evaluations were done in a blinded fashion, independent of

156 the knowledge of the prediction by the DL algorithm as well as to other clinical characteristics
157 like age, sex and comorbidities at the time of admission.
158
159 **Statistical analyses**
160 Descriptive statistics included mean and standard deviation for continuous variables and
161 proportions for categorical variables. Agreement between PCC Experts' evaluation and the DL
162 algorithm's prediction with the ground truth was assessed using Cohen's kappa. Performance
163 metrics for the image classification task were precision (synonymous with positive predictive
164 value as used in epidemiology), recall (synonymous with sensitivity), accuracy and F1 score
165 (which was estimated as the harmonic mean of precision and recall). In addition, area under a
166 receiver operating characteristic curve (AUROC) was estimated for the DL predictions.
167 Predictive performance of the DL model was assessed at the level of the image as well as at the
168 level of the patient. To summarize the performance at the level of a patient, we considered the
169 prediction to be "mechanical ventilation needed" if any of the multiple X-ray images on the
170 same patient had indicated a high likelihood of ventilation need by the DL model.
171 Correspondingly, the maximum probability estimated by the DL model for multiple X-rays on a
172 given patient was considered as the predicted probability of the need for mechanical
173 ventilation at the level of the patient. Prognostic value of the predictions from the deep
174 learning model and evaluations from the PCC experts was conducted using Kaplan-Meier plots
175 and Cox proportional hazards models. Incremental performance attributable to the deep
176 learning model was estimated using the Harrell's C statistic for survival models [14, 15] and
177 compared for statistical significance using the likelihood χ^2 test. All statistical analyses were

178 conducted in Stata 12.0 (Stata Corp, College Station, TX) software package. A global type I error
179 rate of 0.05 was used to test statistical significance.

180
181 **RESULTS**

182

183 **Study participants**

184 Data and images for this study come from 528 COVID-19 positive, hospitalized patients and a
185 total of 663 X-ray images (Figure 1). Of the 528 patients, 79 (~15%) required mechanical
186 ventilation. Clinical characteristics of the study participants based on the need for mechanical
187 ventilation are shown in Table 1. Interestingly, none of the socio-demographic and comorbidity
188 variables were statistically significantly different in patients who received mechanical
189 ventilation as compared to those who did not. However, in general, those who received
190 mechanical ventilation were more likely to be aged over 60 years and have hypertension,
191 obesity, diabetes or chronic kidney disease as a comorbidity. The death rate in those who were
192 mechanically ventilated was very high (~66%) as compared to those who did not need
193 mechanical ventilation (~4%) as shown in Table 1.

194

195 **Model training results**

196 The results of training of the proposed model are shown in Figure 2E. The loss function
197 monotonically (except for cycle length 7) decreased in both the training and the validation
198 subsets and, conversely, the accuracy of prediction increased in a mirror-image fashion in both
199 subsets. The model achieved convergence quickly. At the end of 10 cycle lengths, the training
200 set and test set accuracy was very high – almost 100% in the training set and 100% in the

validation set. As shown in Figure 2F, the model perfectly predicted the need for mechanical ventilation in the validation set.

Predictive performance of the model in the test set

The predictive performance of the model was assessed at two levels – at the level of X-ray images (n = 153) and at the level of an individual patient (n = 118). These results are shown in Figure 3 (panels A-B for image-level analyses and panels E-F for patient-level analyses). The ROC curve using mechanical ventilation needed (22 patients, 43 X-ray images) as the ground truth and the probability estimates from the DL model as predictor (Figure 3A) showed an AUROC of 79.34% at the image-level. The optimum cutoff point on this ROC had a sensitivity and specificity of 70% and 84%, respectively. The confusion matrix (Figure 3B) showed that the performance of the DL model was good with a high accuracy (0.7974), good recall, precision and F1-score (0.6976, 0.6250 and 0.6593) as well as a good Cohen's kappa (0.5158).

We replicated these analyses at the level of the patient with the maximum predicted probability (from multiple X-ray images). We observed (Figure 3E) that the AUROC increased to 90.06% with an optimum sensitivity and specificity of (86.34% and 84.38%, respectively). Comparing these estimates with the corresponding image-level estimates (Figure 3A), we found that analyses at the level of the patient yielded substantially higher sensitivity without loss of specificity. The confusion matrix for comparison of the patient-level prediction with ground truth (Figure 3F) showed a markedly improved predictive performance: accuracy (0.8474), recall (0.8636), precision (0.5588), F1-score (0.6786) and Cohen's kappa (0.5845).

Independent evaluation by PCC Experts

Independent evaluations by the two PCC Experts are shown as confusion matrices for image-level analyses (Figure 3C-D) and patient-level analyses (Figure 3G-H). Performance of both PCC Experts was relatively lower as compared to the DL model at the image-level as well as at the patient-level. At the image-level the performance characteristics of PCC Expert 1 were: accuracy (0.7451), recall (0.6744), precision (0.5370), F1-score (0.5979) and Cohen's kappa (0.4148). Similarly, the performance characteristics of PCC Expert 2 at the level of images were: accuracy (0.7059), recall (0.8140), precision (0.4861), F1-score (0.6087) and Cohen's kappa (0.3962). Like the DL model performance, the performance characteristics improved when the analyses were done at the level of patients (Figure 3G-H). For example, the performance characteristics of PCC Expert 1 at the level of patient were: accuracy (0.8051), recall (0.8636), precision (0.4872), F1-score (0.6230) and Cohen's kappa (0.5049). For PCC Expert 2 the performance characteristics were: accuracy (0.7119), recall (0.9090), precision (0.3846), F1-score (0.5405) and Cohen's kappa (0.3774). Despite these improved estimates at the level of the patient and with the exception of recall for PCC Expert 2, all the performance characteristics of both the PCC Experts were either on par or below the corresponding estimates for the DL model.

Incremental predictive performance of DL model

To assess the incremental predictive performance of the DL model we conducted survival analyses with time to mechanical ventilation as the outcome of interest. These results are shown in Figure 4. Kaplan-Meier plots (Figure 4A-C) showed that patients predicted to need

mechanical ventilation by the DL model or the PCC Experts rapidly progressed to mechanical ventilation (red curves in Figure 4A-C). However, the relative hazards of progressing to mechanical ventilation were highest for the DL model (15.3) as compared to those of PCC Experts 1 (11.3) and PCC Expert 2 (9.9) indirectly implying better prognostic stratification by the DL model. To directly assess the incremental value of the DL model over the stratification done by the PCC Experts, we conducted pairwise comparisons of a series of Cox proportional hazards models using the Harrell's C-statistic. This statistic was estimated to be 0.7454 for stratification offered by PCC Expert 1 but increased to 0.8331 (improvement 0.0877, $p<0.001$) upon addition of DL model prediction as a covariate in the Cox Model (compare models 1 and 1A in Figure 4D). Similarly, the addition of DL model prediction to the stratification offered by PCC Expert 2 improved Harrell's C-statistic from 0.6921 to 0.8246 (improvement 0.1325, $p<0.001$). Lastly, when stratifications offered by both the PCC Experts were simultaneously used as covariates, the Harrel's C-statistic was estimated to be 0.7685 which increased to 0.8382 upon addition of the DL model prediction as a covariate (improvement 0.0724, $p = 0.001$). Together, the results in Figure 4 demonstrate that the DL model significantly and incrementally contributed to an improved prediction of the need for and time to mechanical ventilation over and beyond the predictions obtained from two PCC Experts.

Time gained by early prediction using the DL model

Lastly, we examined the time gained by using the DL model predictions for the need of mechanical ventilation. These analyses were done at the level of the patient with start point defined as the time at which the DL model first predicted the need for mechanical ventilation.

Using this strategy, we observed that the median time to mechanical ventilation was 2.98 (95% CI 1.63 – 4.32) days. Thus, the DL model developed in this study antedated mechanical ventilation early at the time of or during index hospitalization.

DISCUSSION

We have developed a novel, X-ray image-based, deep learning model to predict the need for mechanical ventilation early during hospitalization of COVID-19 patients. Our model was accurate (90% at the level of the patient), externally validated in an independent test set and provided improved prediction as compared to the prognostic performance of stratification provided by two PCC Experts. Considering the urgent need for effective rationalization of health care resources for COVID-19 patients, especially the ventilators, we believe that our DL model can have an important role in critical care of COVID-19 patients. This anticipation is contingent upon the observation that our DL model was able to predict the need for mechanical ventilation approximately 5 days ahead of the actual intubation event.

Previously, the CheXNeXt deep learning system [11] has been used to predict 14 pathologies based on chest X-rays. Also, recently Li et al [16] have developed a deep-learning Siamese network to predict the Radiographic Assessment of Lung Edema (RALE) scores [17] used to quantify severity of ARDS in COVID-19 patients. These landmark studies have proffered definitive directions for the potential use of chest X-rays in clinical care of critical patients. However, a direct application of these systems to predict actionable outcomes like the need for mechanical ventilation is currently lacking. Using cytokine / chemokine data on hospitalized COVID-19 patients, Donlan et al [12] have shown that circulating concentration of interleukin-

13 (IL-13) can predict the need for mechanical ventilation. Since IL-13 can contribute to pulmonary eosinophilia and tissue remodeling, it is conceivable that radiographically detectable texture alterations may accompany these cytokine profiles. This hypothesis is, in part, supported by investigations in COVID-19 negative patients.[18, 19] Whether such a correlation exists within the context of COVID-19 is currently unknown. Considered in totality, these previous studies provide a possible biological explanation to why chest radiographs can predict the need for mechanical ventilation in immediate future.

The results of our study should be considered in the light of some limitations. First, this was a retrospective, observational evaluation and the confounding and bias implicit in such an investigation will remain a limitation. Second, the data for this study were derived from a single center and the generalizability of this approach to other settings needs to be established in further studies. Third, we restricted our model to the use of chest radiographs only. However, additional clinical parameters at the time of hospital admission such as respiratory rate, oxygenation status (e.g. the ROX index[20]) and altered mental status [21] along with socio-demographic characteristics, comorbidity profile and laboratory investigations can potentially further improve the prediction. Future studies need to evaluate these possibilities, but our focus was to use an objective measure such as a chest radiograph and provide a tool to the critical care provider with a reasonable expectation of the future course of disease in a given patient.

CONCLUSIONS

312 Till the time effective preventive and management options for COVID-19 patients become
313 widely available, consorted efforts that reduce the risks to the patient and thus the burden on
314 healthcare system are needed. To that end, our study demonstrates a proof-of-principle that
315 early prediction of the need for ventilation using chest X-ray images acquired early during
316 hospitalization can accurately predict the need for mechanical ventilation in COVID-19 patients.
317 We believe that such a tool has value in effectively triaging COVID-19 patients at the time of
318 initial healthcare contact.

REFERENCES

[1] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), 2020 [accessed April 25.2020].

[2] Coronavirus disease 2019, https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html; 2020 [accessed August 11.2020].

[3] Abate SM, Ahmed Ali S, Mantfardo B, Basu B. Rate of Intensive Care Unit admission and outcomes among patients with coronavirus: A systematic review and Meta-analysis. PloS one 2020;15(7):e0235653.

[4] Current hospital capacity estimates - snapshot, https://www.cdc.gov/nhsn/covid19/report-patient-impact.html; 2020 [accessed August 11.2020].

[5] Duca A, Memaj I, Zanardi F, Preti C, Alesi A, Della Bella L, et al. Severity of respiratory failure and outcome of patients needing a ventilatory support in the Emergency Department during Italian novel coronavirus SARS-CoV2 outbreak: Preliminary data on the role of Helmet CPAP and Non-Invasive Positive Pressure Ventilation. EClinicalMedicine 2020;24:100419.

[6] Zhu W, Wang Y, Xiao K, Zhang H, Tian Y, Clifford SP, et al. Establishing and Managing a Temporary Coronavirus Disease 2019 Specialty Hospital in Wuhan, China. Anesthesiology: The Journal of the American Society of Anesthesiologists 2020;132(6):1339-45.

[7] USNS Mercy, USNS Comfort receiving patients in LA, New York City, [accessed August 11.2020].

[8] Interim clinical guidance for management of patients with confirmed coronavirus disease, 2020 [accessed August 11.2020].

[9] Khan AI, Shah JL, Bhat MM. Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Computer Methods and Programs in Biomedicine 2020:105581.

[10] Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. European Respiratory Journal 2020.

[11] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine 2018;15(11):e1002686.

[12] Donlan AN, Young M, Petri WA, Abhyankar M. IL-13 Predicts the Need for Mechanical Ventilation in COVID-19 Patients. medRxiv : the preprint server for health sciences 2020:2020.06.18.20134353.

[13] Artacho Ruiz R, Artacho Jurado B, Caballero Gueto F, Cano Yuste A, Durban Garcia I, Garcia Delgado F, et al. Predictors of success of high-flow nasal cannula in the treatment of acute hypoxemic respiratory failure. Medicina intensiva 2019.

[14] Alexander M, Wolfe R, Ball D, Conron M, Stirling RG, Solomon B, et al. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. British journal of cancer 2017;117(5):744-51.

[15] Peters M, van der Voort van Zyp JRN, Moerland MA, Hoekstra CJ, van de Pol S, Westendorp H, et al. Development and internal validation of a multivariable prediction

model for biochemical failure after whole-gland salvage iodine-125 prostate brachytherapy for recurrent prostate cancer. Brachytherapy 2016;15(3):296-305.

[16] Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, et al. Automated assessment of COVID-19 pulmonary disease severity on chest radiographs using convolutional Siamese neural networks. medRxiv : the preprint server for health sciences 2020.

[17] Warren MA, Zhao Z, Koyama T, Bastarache JA, Shaver CM, Semler MW, et al. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS. Thorax 2018;73(9):840-6.

[18] Svenningsen S, Haider E, Boylan C, Mukherjee M, Eddy RL, Capaldi DPI, et al. CT and Functional MRI to Evaluate Airway Mucus in Severe Asthma. Chest 2019;155(6):1178-89.

[19] Kim YH, Kim KW, Lee KE, Lee MJ, Kim SK, Kim SH, et al. Transforming growth factor-beta 1 in humidifier disinfectant-associated children's interstitial lung disease. Pediatric pulmonology 2016;51(2):173-82.

[20] Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernandez G, et al. An Index Combining Respiratory Rate and Oxygenation to Predict Outcome of Nasal High-Flow Therapy. American journal of respiratory and critical care medicine 2019;199(11):1368-76.

[21] Lee JY, Kim HA, Huh K, Hyun M, Rhee JY, Jang S, et al. Risk Factors for Mortality and Respiratory Support in Elderly Patients Hospitalized with COVID-19 in Korea. Journal of Korean medical science 2020;35(23):e223.

385 **Table 1. Baseline characteristics of study participants (n=528).** Cells indicate the number (N)

386 and proportion (%)

387

Characteristic†	MV Needed (N = 79)	MV not needed (N = 449)	P
Socio-demographic characteristics			
Age >60y	36 (46.57)	163 (36.30)	0.117
Males	51 (64.56)	307 (68.37)	0.503
Hispanic/Latino ethnicity	44 (55.70)	259 (57.68)	0.742
Black / African American race	31 (39.24)	155 (34.52)	0.418
Comorbidities			
Hypertension	36 (46.75)	170 (37.95)	0.144
Obesity	42 (53.16)	195 (43.43)	0.109
Diabetes	40 (51.95)	183 (40.85)	0.069
Coronary artery disease	5 (6.49)	38 (8.48)	0.659*
Chronic kidney disease	9 (11.69)	27 (6.03)	0.085*
Asthma	3 (3.90)	36 (8.04)	0.246*
Chronic liver disease	7 (9.09)	23 (5.13)	0.182*
Congestive heart failure	2 (2.60)	23 (5.13)	0.560*
COPD	4 (5.19)	18 (4.02)	0.548*
ESRD	5 (6.49)	16 (3.57)	0.214*

HIV/AIDS	3 (3.90)	14 (3.13)	0.726*
Atrial fibrillation	1 (1.30)	20 (4.46)	0.340*
Ever smoker	19 (24.05)	87 (19.38)	0.339
Outcomes			
Death	52 (65.82)	17 (3.79)	<0.001

388 *, Fisher's exact test

389 **FIGURE LEGENDS:**

390 **Figure 1: Overall analysis pipeline.** P, number of patients; X, number of X-ray images

391

392 **Figure 2: Densenet121 model, data preprocessing and model training.** (A) Example of a
393 preprocessed X-ray image submitted to modeling. (B) The DenseNet121 architecture.
394 Convolutional layers are prefixed with C (cyan), dense blocks with D (black) and transition
395 blocks (orange) with T. GAP, MP, SM and Sigmoid indicate the global average pooling,
396 maxpooling, softmax and binarization layers within the classifier portion of DenseNet121. Inset
397 shows a dense block with 4 layers and depicts how each succeeding layer receives inputs from
398 all preceding layers. Shown within each proportionately sized colored block is the output size in
399 pixels. (C-D) Data preprocessing. Shown in panel C is a batch of resized X-ray images. Panel D
400 shows the same batch after data augmentation that included center cropping, rotation and
401 horizontal displacement. (E) Training log of DenseNet121 to predict the need for mechanical
402 ventilation. Left axis shows the categorical cross-entropy loss at the end of each cycle length
403 and the right axis shows the estimated accuracy of prediction. Results are shown separately for
404 the training (n = 2142) and the validation (n=378) set of X-ray images. (F) Confusion matrix at
405 the end of DenseNet121 training. All the images were correctly classified at this stage.

406

407 **Figure 3: Prediction for the need of mechanical ventilation.** Analyses were done at the level of
408 X-ray image (A-D) and at the level of each patient (E-H). Panels A and E show the predictive
409 accuracy as AUROC. The optimum cutoff was chosen as the point on ROC closest to the upper
410 left corner of the plot and is indicated by a color-coded circle. The sensitivity (dashed

411 perpendicular to y axis) and specificity (inverse of the dashed perpendicular to the x axis) at the
412 optimal cutoff is shown as Sn(best) and Sp(best), respectively. AUROC, area under the receiver
413 operating characteristic curve; CI, confidence interval. (B-D) Each panel shows the confusion
414 matrix on the left side and five performance metrics in a bar chart on the right side. The bars
415 and error bars show the point and 95% confidence interval for each indicated and (color-coded)
416 performance metric. The metrics shown in the plot are: P, precision; R, recall; A, accuracy; K,
417 Cohen's kappa; and F, F1 score. (F-H) These panels respectively correspond to B-D but the
418 results are shown at the level of the patient. Panels B-D and panels F-H use the same horizontal
419 scale.
420
421 **Figure 4. Incremental prognostic value of the DL model as compared to the PCC Experts'**
422 **evaluation.** (A-C) Kaplan-Meier plots for time to mechanical ventilation since the time of first X-
423 ray image. For patients with multiple X-ray images the time was left-censored at the first image
424 indicating the need of mechanical ventilation. Panels A-C indicate classifications based on the
425 DL model (A), PCC Expert 1 (B) and PCC Expert 2 (C), respectively. Relative hazards (RH) and 95%
426 confidence intervals (CI) were estimated using Cox proportional hazards (PH) models. Since
427 different patients were classified as needing mechanical ventilation (MV) by the DL model and
428 the PCC Experts, different shades of red (for MV needed) and blue (for MV not needed) are
429 used. (D) Incremental value of DL model to prognosticate patients. Models 1 and 1A compare
430 the prediction from a Cox PH model that used only PCC Expert 1 (Model 1) versus that from a
431 Cox PH model that used PCC Expert 1 and the DL model as covariates. Models 2 and 2A,
432 correspondingly compare models with only PCC Expert 2 and PCC Expert 2 with DL model as

433 covariates. Models 3 and 3A compare models with PCC Experts 1 and 2 as covariates and both
434 PCC Experts with DL model, respectively. Bars indicate Harrell's C statistic for the indicated
435 model. The statistical significance for the difference was tested using likelihood χ^2 test and is
436 shown at the top of the bars depicting the indicated paired comparisons.