THE LANCET
Digital Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Jiao Z, Choi JW, Halsey K, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *Lancet Digit Health* 2021; published online March 24. https://doi.org/10.1016/S2589-7500(21)00039-X.

Supplementary Material

Prognostication of COVID-19 patients presenting to the emergency department using artificial intelligence based on chest radiographs and clinical data

Jiao Z, Choi JW, Halsey K, et al

Table of Contents

Appendix A: Additional Materials and Methods

Study Population

In this study, hospitals affiliated with University of Pennsylvania and Brown University were included. From University of Pennsylvania, Hospital of the University of Pennsylvania, Penn Presbyterian Medical Center, Pennsylvania Hospital, and Chester County Hospital were included. From Brown University, Rhode Island Hospital, The Miriam Hospital, Newport Hospital, and Woman & Infants Hospital were included. A diagram illustrating patient inclusion and exclusion criteria is shown in Figure S1.

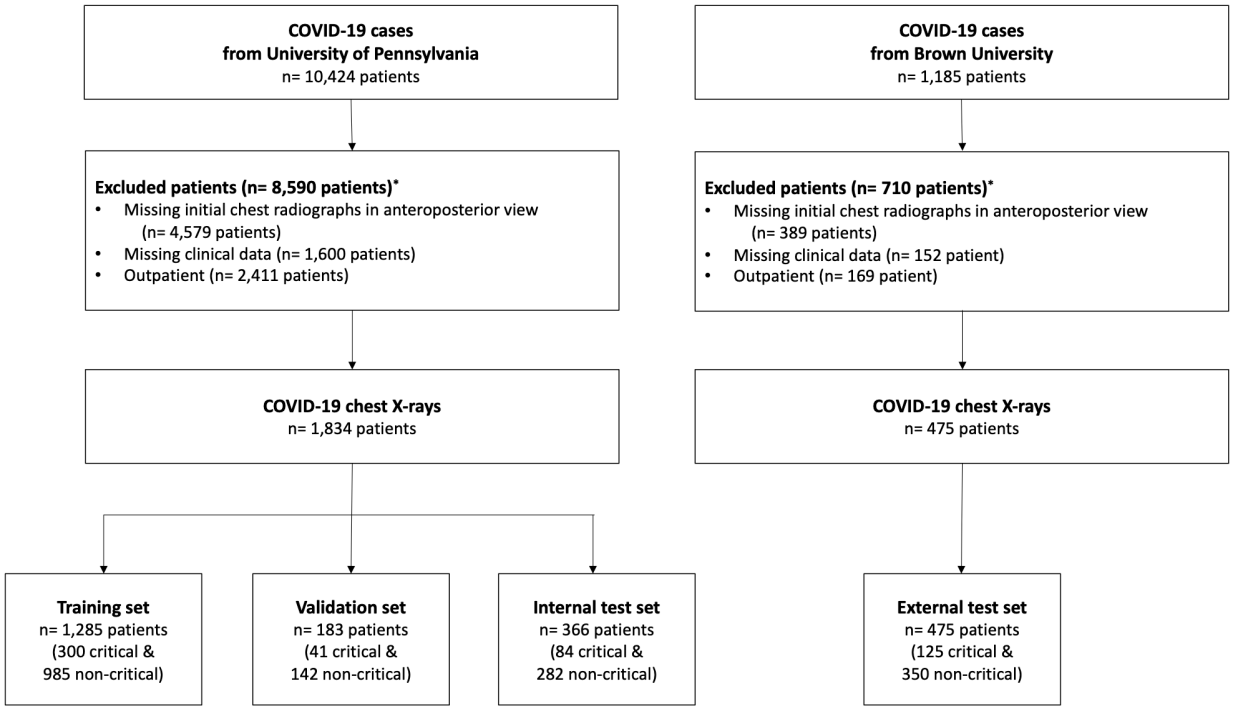

Figure S1. Flow diagram of patient inclusion and exclusion criteria.

CXR Severity Score

The radiologists were blinded to patient information other than COVID-19 positivity. Each CXR was divided into six lung zones (right upper, right middle, right lower, left upper, left middle, and left lower zones), and each lung zone was assigned a score of 0 (no opacity) or 1 (opacity).[1] The final severity score for each lung zone was labeled with a score of 1 only when both radiologists agreed to its opacity. If there was a discrepancy between the two junior radiologists' scoring, then it was resolved in consensus with the senior radiologist. Otherwise, the lung zones were labeled with a score of 0. The final score in all six lung zones were summed to generate the total severity score. Examples of CXR severity scoring are shown in Figure S2.

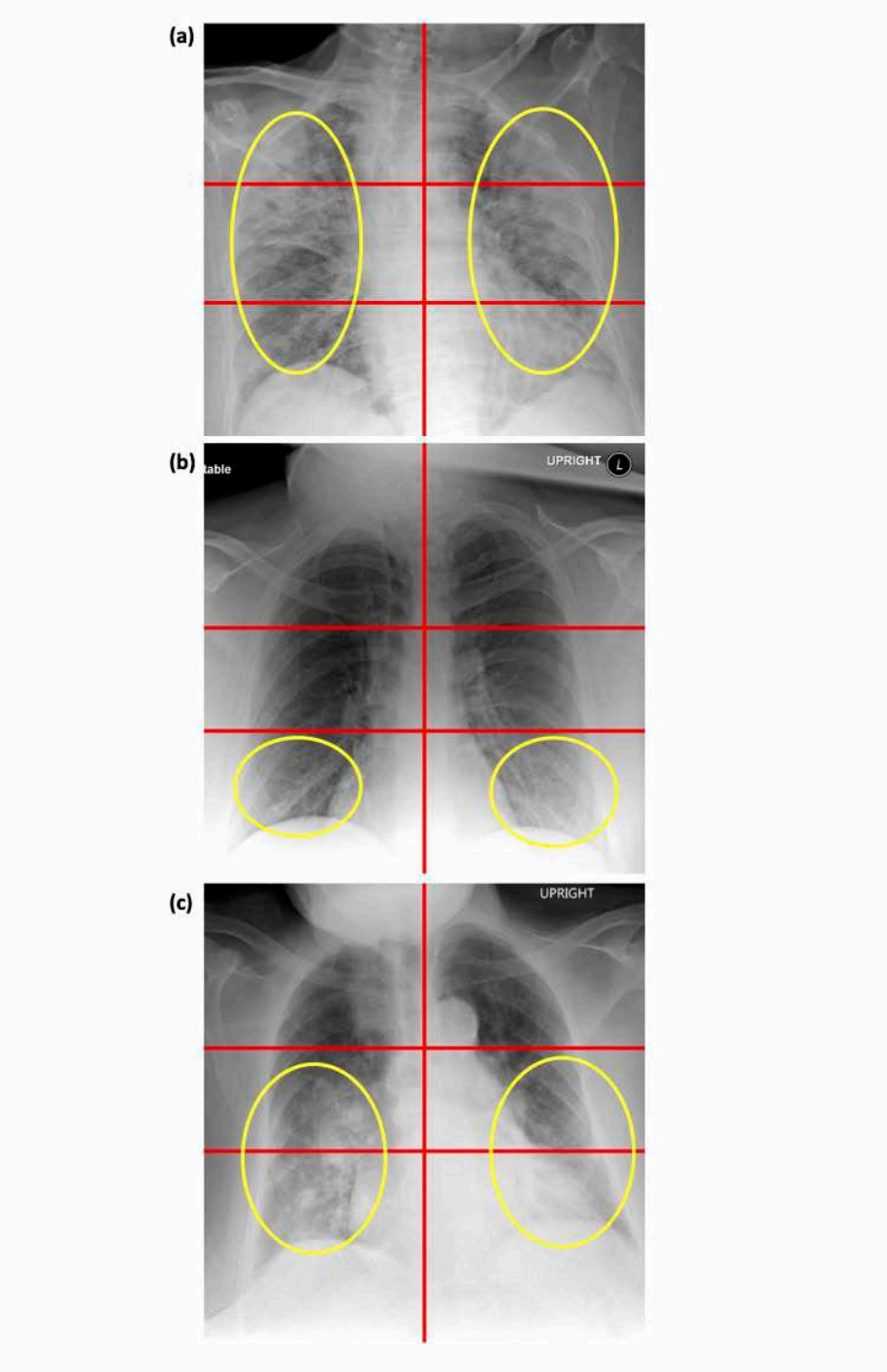

Figure S2. Examples of chest x-ray severity scoring of three patients with COVID-19. (a) Chest x-ray shows hazy opacities in all six right and left lung zones (severity score= 6). (b) Chest x-ray shows hazy opacities in right lower and left lower lung zones (severity score= 2). (c) Chest x-ray shows hazy opacities in left middle, left lower, right middle, and right lower lung zones (severity score= 4).

CXR Segmentation

The automatic segmentations of the lung parenchyma were manually annotated by medical students (KH, LT, BH, JC) followed by edits from a board-certified radiologist (HXB). For any incorrect automatic segmentations, adjustments were manually made using the open source 3D Slicer software (version 4.6).[2] A total of 143 out of 2,309 lung segmentations (4%) were manually corrected by a radiology technologist (SC) with confirmation by a board-certified radiologist (HXB). Figure S3 shows an example of auto-segmentation and incorrect auto-segmentation with manual corrections.

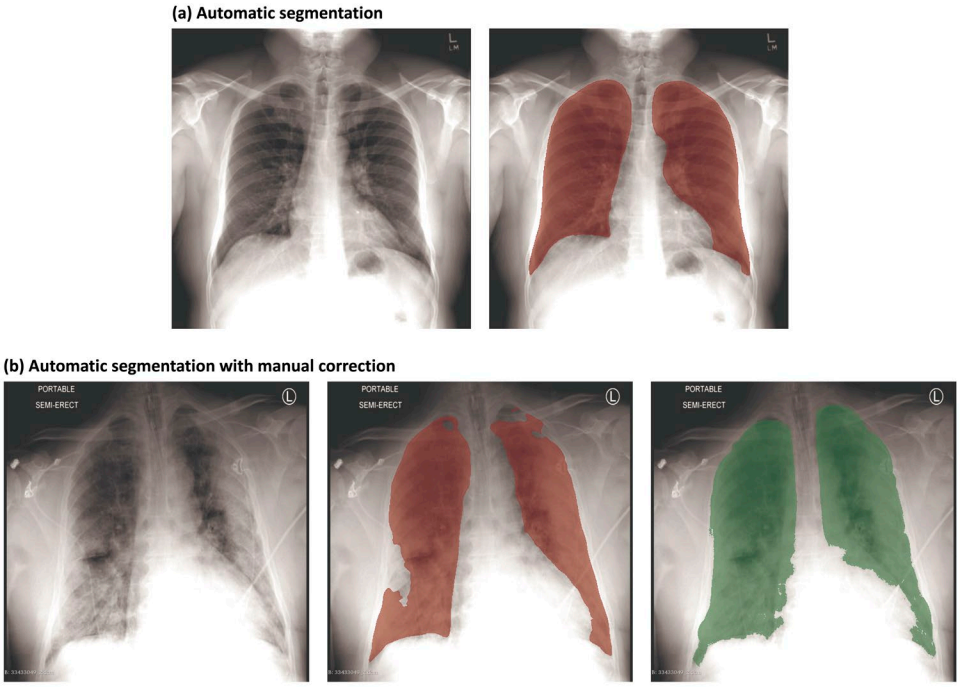

Figure S3. Chest x-rays with automatic segmentation (red) and manual correction (green) of the lungs. (a) Successful automatic segmentation, which did not require manual corrections. (b) Incorrect automatic segmentation with manual corrections.

Severity Prediction Model

The pretrained EfficientNet layers extracted a tensor with the scale of 1280 × 16 × 16 (number of channels × height × width of feature map) from the mask wrapped lung region of the CXR. Then, the feature representation was forwarded through one convolutional layer (256) with global average pooling operation and three additional dense layers (256, 32, 2). These specific parameters were used because they achieved the best performance on validation in comparison to other parameters (Figure S4).

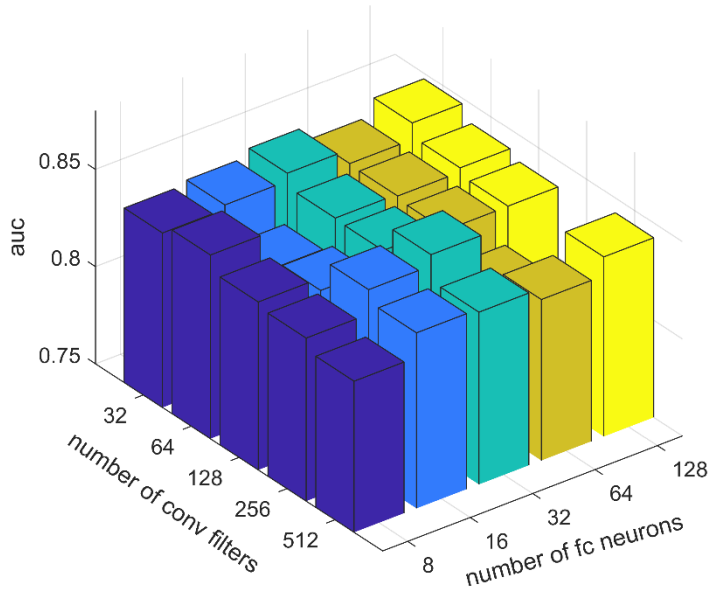

Figure S4. Performance of the deep learning model in different parameters on validation set. The imaging data are passed through feature extraction layer (EfficientNet) and then four severity prediction layers (convolution and dense layers). The network parameters that can be adjusted are the number of filters in the added convolution (conv) layer and number of fc neurons in the penultimate dense layer. Different combinations of these network parameters were evaluated on validation set. When the number of conv filters was set as 256 and the number of fc neurons was set as 32, the model achieved the best validation performance.

Progression Prediction Model

The survival forest is a variant of the random forest adapted for right-censored survival and recurrence data.[3] It is optimized by assigning risk scores to patients according to their critical label and time on the training set. In each of the two survival forest models, a collection of decision trees was used to model the complex relationships between input feature vectors and the rank time to event risk prediction. Specifically, a patient who progressed early to critical outcome was assigned a higher risk score than patients who progressed later. In addition, these models provide a series of cumulative hazard outputs that predict patients' cumulative hazard of progressing to critical illness at each time point.

Appendix B: Additional Results

Patient Cohort

The clinical characteristics of patients in training, validation, internal testing, and external testing sets are shown in Table S1. The clinical characteristics of critical and non-critical patients are shown in Table S2. The distribution of time from CXR to critical event among critical patients is shown in Figure S5.

	Training (n=1,285)	Validation (n=183)	Internal Test (n=366)	External Test (n=475)	p-value
Age (years)					<0.001
Median (IQR)	55 (30)	57 (36.5)	52 (31)	60 (26.5)	
<20	18 (1)	1 (1)	9 (2)	6 (1)	
20-39	340 (26)	54 (30)	99 (27)	66 (14)	
40-59	377 (29)	42 (23)	119 (33)	160 (34)	
60-79	409 (32)	60 (33)	108 (30)	168 (35)	
≥80	140 (11)	26 (14)	31 (8)	75 (16)	
Sex					<0.001
Male	588 (46)	82 (45)	184 (50)	278 (59)	
Female	697 (54)	101 (55)	182 (50)	197 (41)	
Body Temperature (°C)					0.846
Elevated (>37)	822 (64)	123 (67)	241 (66)	311 (65)	
Not elevated (≤37)	450 (35)	59 (32)	123 (34)	164 (35)	
SpO_2 (%)					<0.001
Not decreased (≥ 94)	1056 (82)	149 (81)	300 (82)	345 (73)	
Decreased (<94)	198 (15)	30 (16)	55 (15)	118 (25)	
White Blood Cell Count ($x10^9$/L)					0.009
Elevated (>11)	167 (13)	28 (15)	45 (12)	100 (21)	
Not elevated (≤11)	933 (73)	131 (72)	275 (75)	363 (76)	
Lymphocyte Count ($x10^9$/L)					<0.001
Not decreased (≥1.0)	641 (50)	91 (50)	182 (50)	195 (41)	
Decreased (<1.0)	449 (35)	68 (37)	133 (36)	268 (56)	
Creatinine (mg/dL)					0.014
Elevated (≥1.27)	348 (27)	46 (25)	87 (24)	113 (24)	
Not elevated (<1.27)	728 (57)	110 (60)	224 (61)	353 (74)	
CRP (mg/dL)					0.351
Elevated (≥1.0)	295 (23)	46 (25)	84 (23)	299 (63)	
Not elevated (<1.0)	27 (2)	7 (4)	7 (2)	40 (8)	
Comorbidities					
Cardiovascular Disease	287 (22)	37 (20)	66 (18)	124 (26)	0.037
Hypertension	493 (38)	60 (33)	129 (35)	201 (42)	0.058
COPD	68 (5)	14 (8)	8 (2)	32 (7)	0.011
Diabetes	284 (22)	34 (19)	77 (21)	114 (24)	0.447
Chronic Liver Disease	37 (3)	4 (2)	9 (2)	12 (3)	0.925
Chronic Kidney Disease	158 (12)	20 (11)	37 (10)	40 (8)	0.133
Malignant Tumor	67 (5)	11 (6)	14 (4)	24 (5)	0.665

HIV	21 (2)	1 (1)	5 (1)	9 (2)	0.632
Severity					0.543
Critical	300 (23)	41 (22)	84 (23)	125 (26)	
Non-critical	985 (77)	142 (78)	282 (77)	350 (74)	
Outcomes					
Inpatient admission*	756 (59)	108 (59)	218 (60)	412 (87)	<0.001
ICU admission	248 (19)	35 (19)	77 (21)	92 (19)	0.898
Mechanical ventilator	167 (13)	21 (11)	55 (15)	70 (15)	0.520
Death	94 (7)	14 (8)	30 (8)	51 (11)	0.136
Discharged	1,186 (92)	169 (92)	335 (92)	412 (87)	0.003
Progression from Chest X-Ray to Critical Event					0.804
Median (IQR)	0.57 (2.44)	0.63 (2.45)	0.63 (2.42)	0.76 (2.91)	
Day 1	155 (12)	22 (12)	44 (12)	62 (13)	
Day 2	36 (3)	5 (3)	10 (3)	13 (3)	
Day 3	19 (1)	3 (2)	6 (2)	11 (2)	
>Day 3	61 (5)	9 (5)	17 (5)	33 (7)	
Censored**	37 (3)	5 (3)	10 (3)	6 (1)	

Table S1. Comparison of patient characteristics across training and validation, internal test, and external test sets. Abbreviations: SpO2- oxygen saturation on room air, CRP- c-reactive protein, IQR- interquartile range, COPD- chronic obstructive pulmonary disease, HIV- human immunodeficiency virus, and ICU- intensive critical unit.
*Inpatient admission includes ICU admission.
**Censored includes patients whose chest x-ray and clinical data were taken during or after a critical event.

	Critical (n=550)	Non-Critical (n=1,759)	p-value
Age (year)			<0.001
Median (IQR)	67 (22)	51 (29)	
<20	4 (1)	30 (2)	
20-39	49 (9)	510 (29)	
40-59	120 (22)	579 (33)	
60-79	262 (48)	483 (27)	
≥80	115 (21)	157 (9)	
Sex			<0.001
Male	306 (56)	826 (47)	
Female	244 (44)	933 (53)	
Body Temperature (ºC)			<0.001
Elevated (>37)	421 (77)	1,076 (61)	
Not elevated (≤37)	126 (23)	670 (38)	
SpO_2 (%)			<0.001
Not decreased (≥ 94)	326 (59)	1524 (87)	
Decreased (<94)	207 (38)	194 (11)	
White Blood Cell Count ($x10^9$/L)			<0.001
Elevated (>11)	192 (35)	148 (8)	
Not elevated (≤11)	357 (65)	1,345 (76)	
Lymphocyte Count ($x10^9$/L)			<0.001
Not decreased (≥1.0)	209 (38)	900 (51)	
Decreased (<1.0)	338 (61)	580 (33)	
Creatinine Level (mg/dL)			<0.001
Elevated (≥1.27)	262 (48)	332 (19)	
Not elevated (<1.27)	280 (51)	1,135 (65)	
CRP (mg/dL)			<0.001
Elevated (≥1.0)	286 (52)	438 (25)	
Not elevated (<1.0)	9 (2)	72 (4)	
Comorbidities			
Cardiovascular Disease	228 (41)	286 (16)	<0.001
Hypertension	308 (56)	575 (33)	<0.001
COPD	64 (12)	58 (3)	<0.001
Diabetes	186 (34)	323 (18)	<0.001
Chronic Liver Disease	25 (5)	37 (2)	0.002
Chronic Kidney Disease	116 (21)	139 (8)	<0.001
Malignant Tumor	52 (9)	64 (4)	<0.001
HIV	8 (1)	28 (2)	0.823

Outcomes			
Inpatient admission*	547 (99)	947 (54)	<0.001
ICU admission	452 (82)	N/A	
Mechanical ventilator	313 (57)	N/A	
Death	189 (34)	N/A	
Discharged	354 (64)	1748 (99)	<0.001
Progression from Chest X-Ray to Critical Event			
Median (IQR)	0.63 (2.61)	N/A	
Day 1	283 (51)	N/A	
Day 2	64 (12)	N/A	
Day 3	39 (7)	N/A	
>Day 3	120 (22)	N/A	
Censored**	58 (11)	N/A	

Table S2. Characteristics of critical and non-critical patients with COVID-19. Abbreviations: SpO_2- oxygen saturation on room air, CRP- c-reactive protein, IQR- interquartile range, COPD- chronic obstructive pulmonary disease, HIV- human immunodeficiency virus, and ICU- intensive critical unit.
*Inpatient admission includes ICU admission.
**Censored includes patients whose chest x-ray and clinical data were taken during or after a critical event.

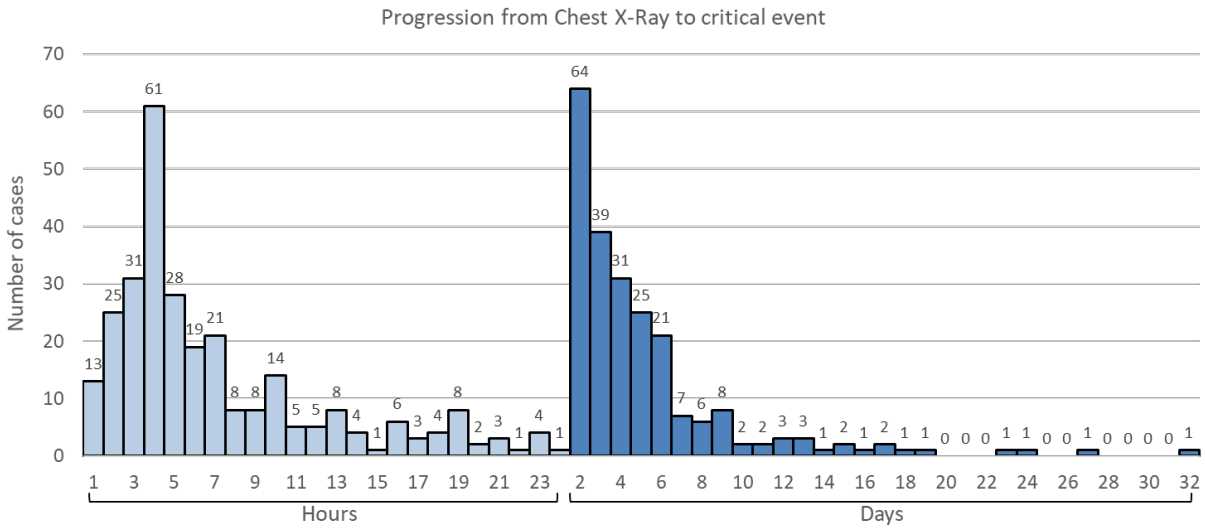

Figure S5. Distribution of time from chest x-ray to critical event. A critical event is defined as utilization of mechanical ventilation, admission to intensive care unit, and/or death among patients with COVID-19.

Severity Prediction Model

The contributions of deep learning features extracted from CXR[4] and clinical variables[5] to the severity prediction model were analyzed. The contribution of deep learning features is visualized in the attention maps (Figure S6). It demonstrates the ability to focus on pulmonary tissue. The contribution of clinical variables is in Figure S7. The Precision-Recall (P-R) curves of the severity prediction model are shown in Figure S8.

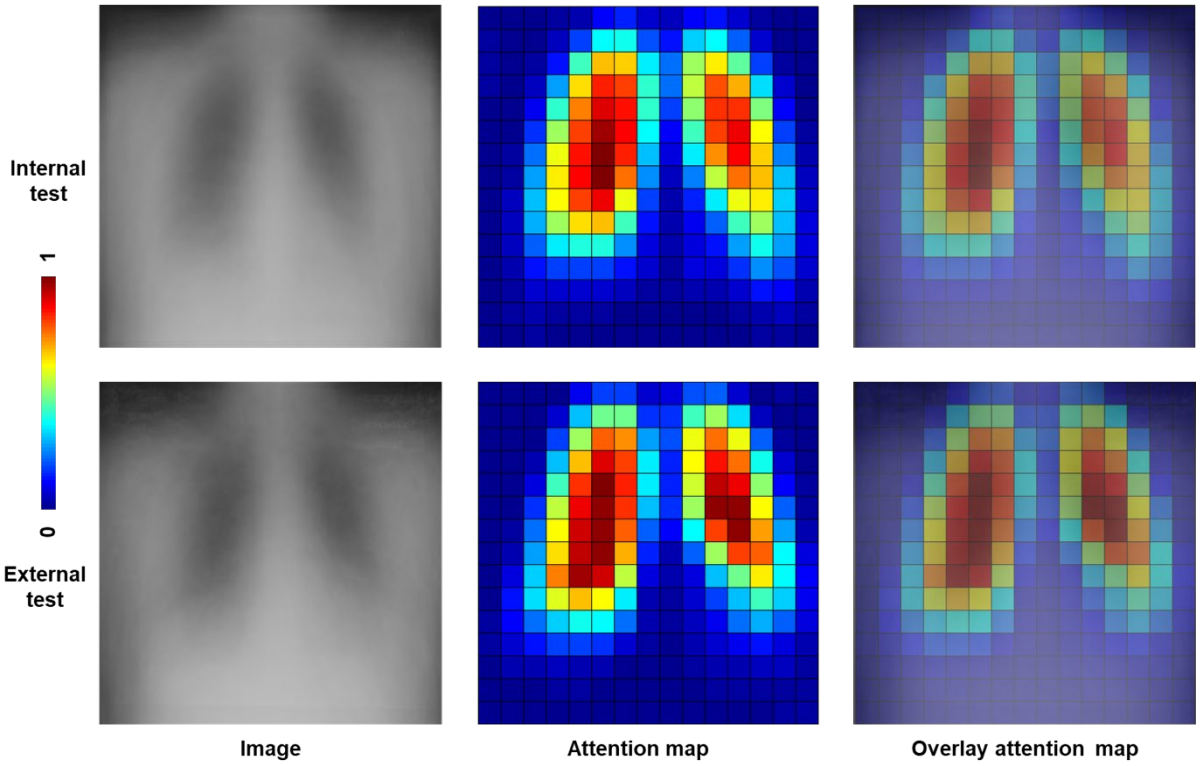

Figure S6. Contribution of deep learning features from chest x-rays to the severity prediction on internal and external test sets. For each test set, the averaged attention map is shown with the averaged image and their overlay. The attention map illustrates the contribution of each image location to severity prediction. The map is in a 16*16 grid like the feature map extracted from EfficientNet and built from calculating the class activation values from each location on the chest x-ray.

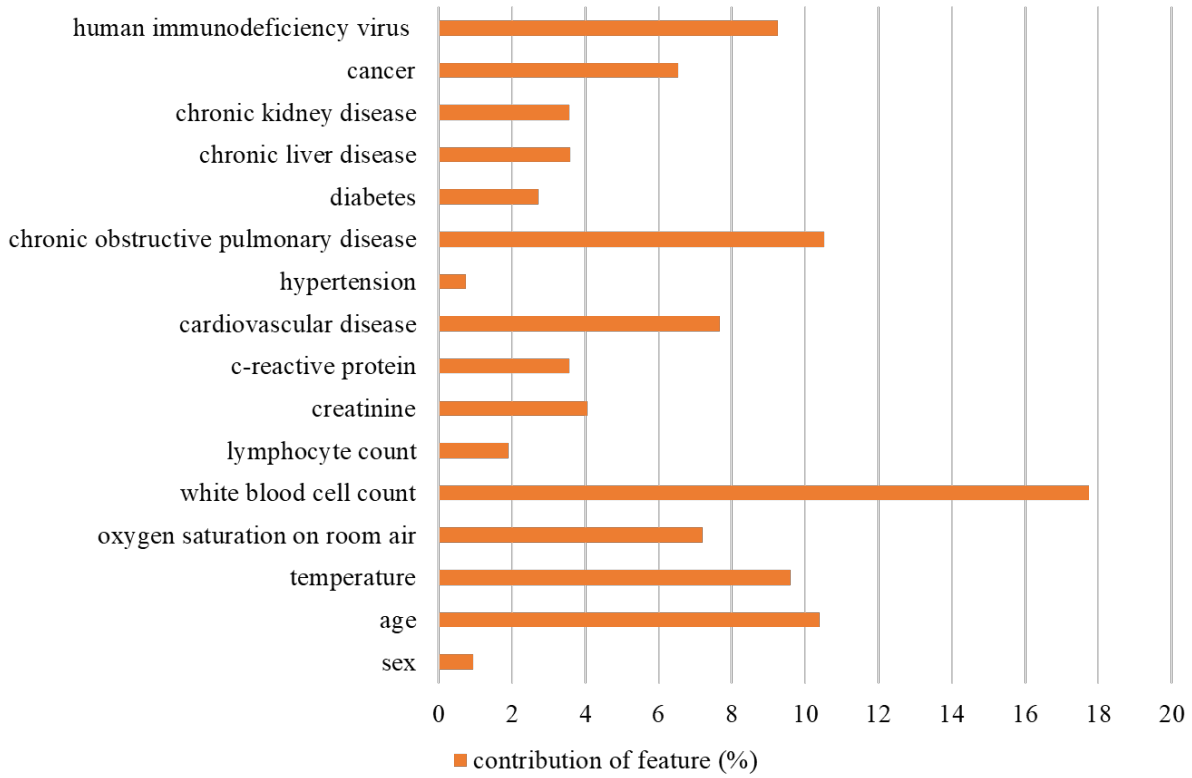

Figure S7. Contribution (in percentage) of each clinical variable to the severity prediction model based on clinical data. A total of sixteen clinical variables were included in this study- patient's age and sex, vital signs (temperature, SpO2 on room air) and lab values (white blood cell count, lymphocyte count, creatinine, and c-reactive protein) at the initial time of presentation, and comorbidities (cardiovascular disease, hypertension, chronic obstructive pulmonary disease, diabetes, chronic liver disease, chronic kidney disease, cancer, and human immunodeficiency virus).

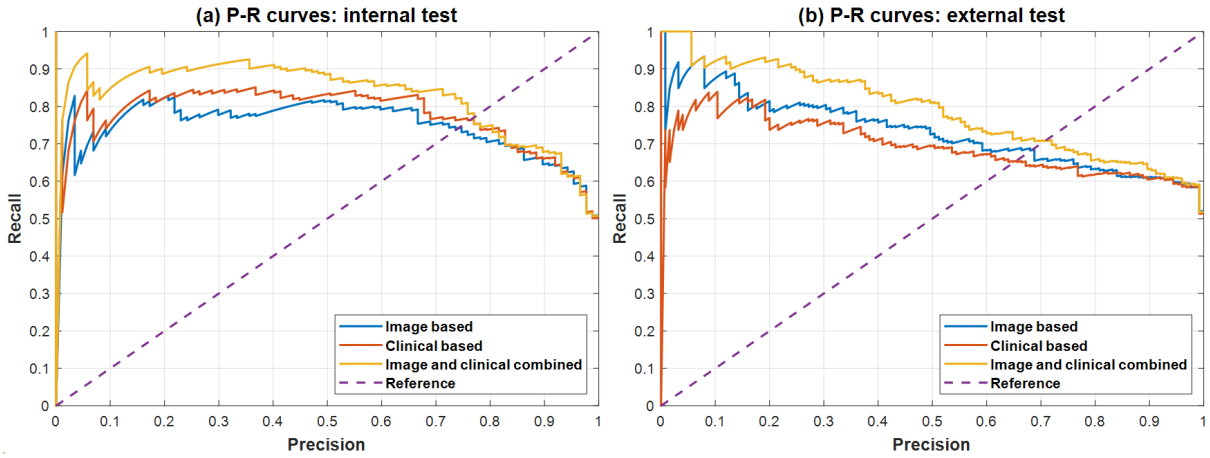

Figure S8. The Precision-Recall curves of severity prediction model on internal and external test sets. The severity prediction model based on image and clinical data obtained the best performance in comparison to the model based on image alone and the model based on clinical data alone on both test sets

Progression Prediction Model

The contribution of deep learning features is visualized in the attention maps (Figure S9). It demonstrates the ability to focus on pulmonary tissue. The contribution of clinical variables is in Figure S10. The P-R curves of the progression prediction model are shown in Figure S11. The cumulative hazards of two stratified risk groups in progressing to a critical outcome at every time point is shown in Figure S12.

In addition, the performance of progression prediction models from CXR to 1) mechanical ventilation and/or ICU admission and 2) death separately was analyzed. They were built and evaluated using the same method and metrics as the progression prediction model from CXR to critical outcome.

The performance of a progression prediction model from CXR to mechanical ventilation and/or ICU admission is summarized in Table S3. The image based model achieved a C-index of 0·726 on the internal test set and a C ·index of 0·724 on the external test set. The clinical data based model achieved a C-index of 0·751 on the internal test set and 0·700 on the external test set. When the deep learning features based prediction were combined with clinical data based analysis, the prediction performance was improved to a C-index of 0·789 on the internal test set and 0·752 on the external test set. The combined progression prediction model had a statistically significant improvement in comparison to the image based prediction and clinical data based prediction. With the same risk stratification method applied in critical prediction task, the two stratified risk groups obtained from all our methods show significant differences. Particularly, the image and clinical combined model achieved the best performance according to log-rank test ($p < 0.0001$, $\chi^2 = 27.08$ on internal testing; and $p < 0.0001$, $\chi^2 = 40.35$ on external testing).

The performance of a progression prediction model from CXR to death is summarized in Table S4. The image based model achieved a C-index of 0·796 on the internal test set and a C ·index of 0·668 on the external test set. The clinical data based model achieved a C-index of 0·750 on the internal test set and 0·732 on the external test set. When the deep learning features based prediction were combined with clinical data based analysis, the prediction performance was improved to a C-index of 0·835 on the internal test set and 0·767 on the external test set. The combined progression prediction model had a statistically significant improvement in comparison to the image based prediction and clinical data based prediction. With the same risk stratification method applied in critical prediction task, the two stratified risk groups obtained from all our methods show significant differences. Particularly, the clinical based model achieved the best performance according to log-rank test ($p = 0.003$, $\chi^2 = 9.05$ on internal testing; and $p < 0.0001$, $\chi^2 = 16.61$ on external testing).

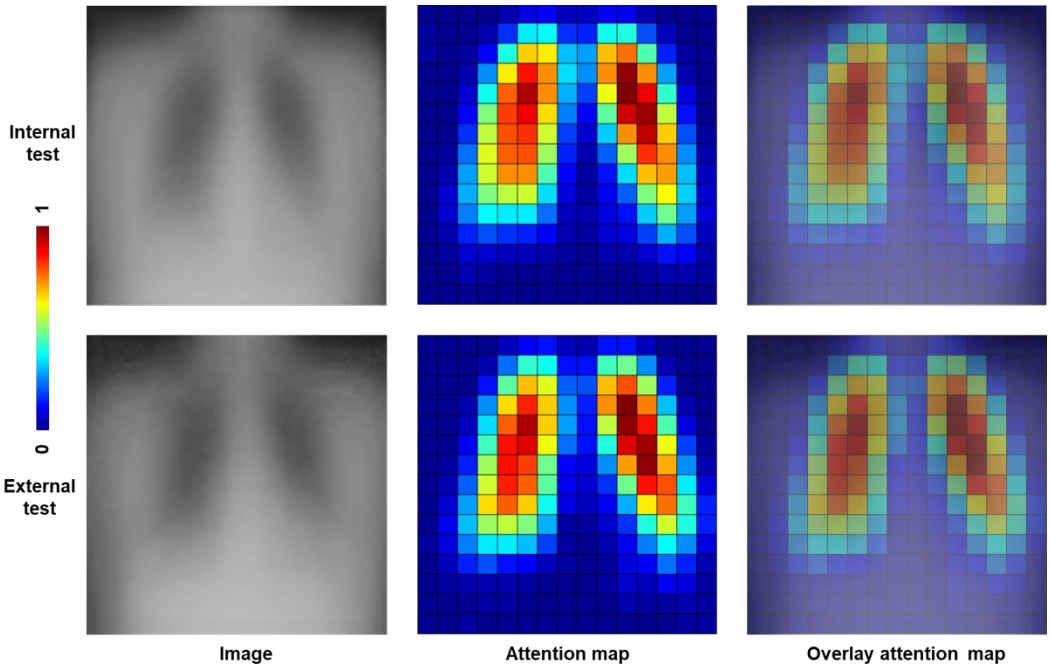

Figure S9. Contribution of deep learning features from chest x-rays to the progression prediction on internal and external test sets. For each test set, the averaged attention map is shown with the averaged image and their overlay. The attention map illustrates the contribution of each image location to progression prediction. The map is in a 16*16 grid like the feature map extracted from EfficientNet and built from calculating the class activation values from each location on the chest x-ray.

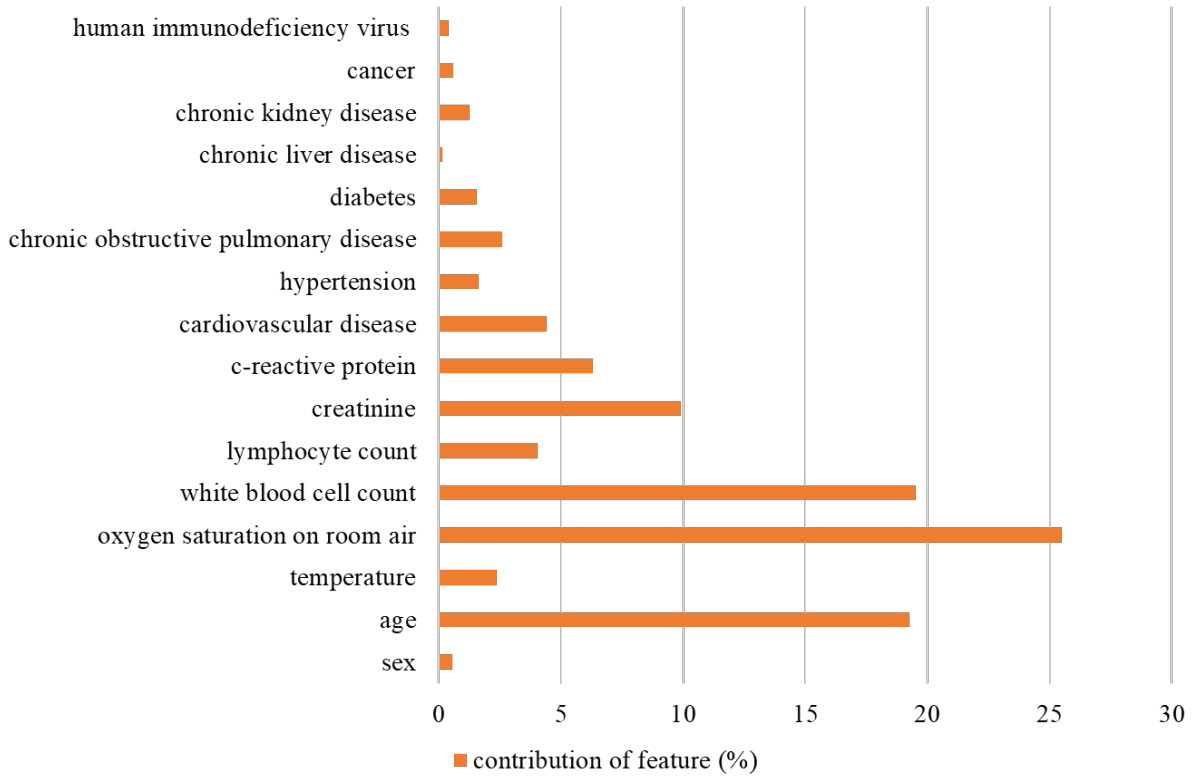

Figure S10. Contribution (in percentage) of each clinical variable to the progression prediction model based on clinical data. A total of sixteen clinical variables were included in this study- patient's age and sex, vital signs (temperature, SpO2 on room air) and lab values (white blood cell count, lymphocyte count, creatinine, and c-reactive protein) at the initial time of presentation, and comorbidities (cardiovascular disease, hypertension, chronic obstructive pulmonary disease, diabetes, chronic liver disease, chronic kidney disease, cancer, and human immunodeficiency virus).

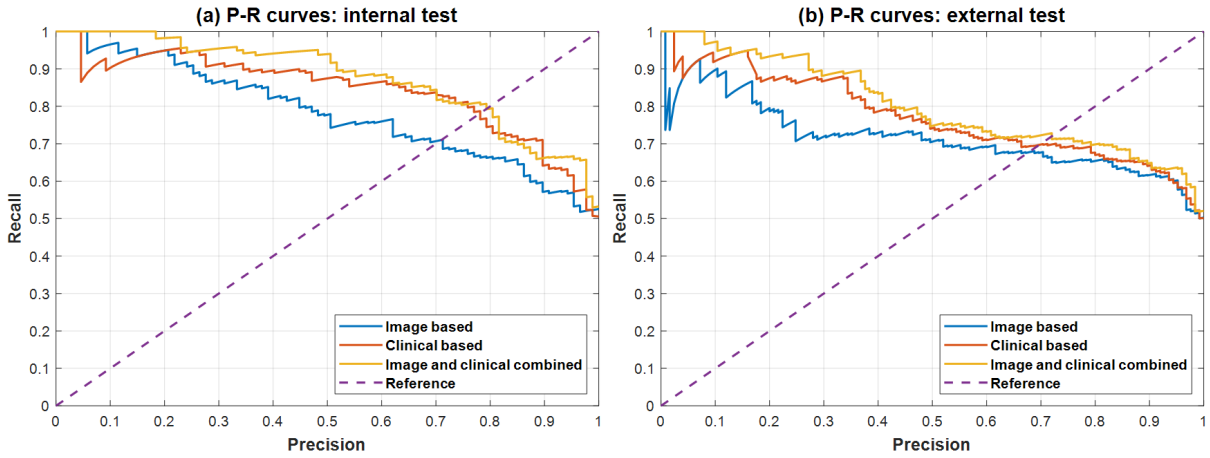

Figure S11. The Precision-Recall curves of progression prediction model on internal and external test sets. The progression prediction model based on image and clinical data obtained the best performance in comparison to the model based on image alone and the model based on clinical data alone on both test sets

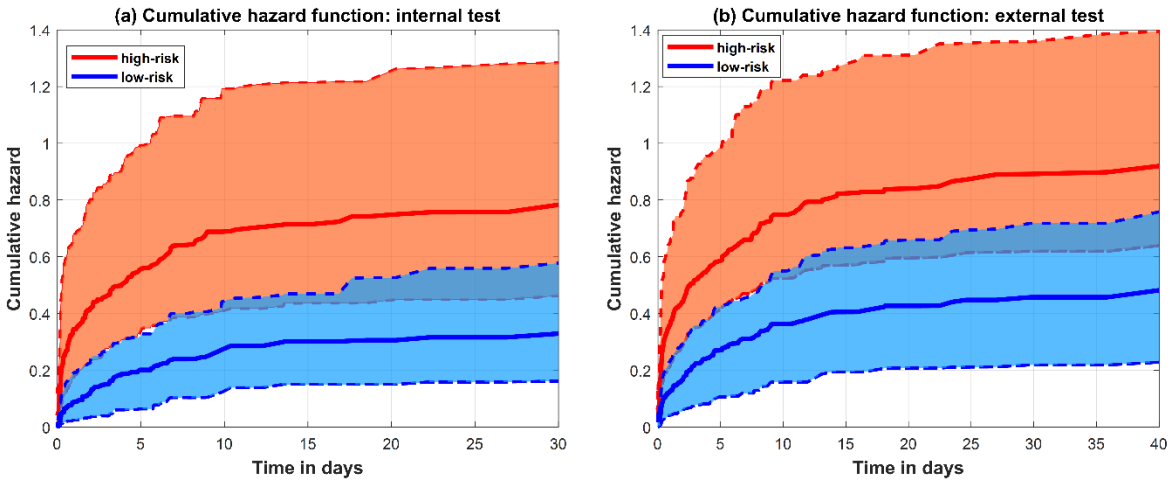

Figure S12. The cumulative hazard functions of two stratified risk subgroups (high- and low-risk) using the combined progression prediction model on internal and external test sets. The solid line is the mean cumulative hazard scores at each time point. The area between the two dashed lines of same color represents the 95% confidence interval.

Progression Prediction Model	Internal Test Set				External Test Set			
	C-index (CI)	p-value*	p-value**	χ^2**	C-index (CI)	p-value*	p-value**	χ^2**
Image based	0.726 (0.700, 0.742)	<0.0001	< 0.0001	19.10	0.724 (0.715, 0.731)	<0.0001	< 0.0001	28.38
Clinical based	0.751 (0.735, 0.784)	<0.0001	< 0.0001	21.35	0.700 (0.688, 0.722)	<0.0001	< 0.0001	25.76
Image and clinical combined	0.789 (0.769, 0.814)	N/A	< 0.0001	27.08	0.752 (0.743, 0.765)	N/A	< 0.0001	40.35

Table S3. Performance of progression prediction model to mechanical ventilation and/or ICU admission based on imaging, clinical data, and severity score on internal and external test sets. Concordance index (C-index) for right-censored data and 95% confidence intervals (CI) measure the model performance by comparing the progression information (critical labels and progression days) with predicted risk scores. A larger C-index correlates with better progression prediction performance.
*A p-value from binomial test measures the difference in performance between the image and clinical combined model and other prediction models. A smaller p-value represents more significant difference between the combined model and other models.
**A p-value from log-rank test between high-risk and low-risk groups and Chi square values (χ^2) show a comparison of stratification performance of different models. A smaller p-value and larger χ^2 correlate with better risk stratification performance.

Progression Prediction Model	Internal Test Set				External Test Set			
	C-index (CI)	p-value*	p-value**	χ^{2}**	C-index (CI)	p-value*	p-value**	χ^{2}**
Image based	0.796 (0.764, 0.823)	<0.0001	0.016	5.77	0.668 (0.610, 0.685)	<0.0001	0.04	4.22
Clinical based	0.750 (0.712, 0.766)	<0.0001	0.003	9.05	0.732 (0.708, 0.749)	<0.0001	< 0.0001	16.61
Image and clinical combined	0.835 (0.815, 0.852)	N/A	0.020	5.44	0.767 (0.740, 0.787)	N/A	< 0.0001	16.29

Table S4. Performance of progression prediction model to death based on imaging, clinical data, and severity score on internal and external test sets. Concordance index (C-index) for right-censored data and 95% confidence intervals (CI) measure the model performance by comparing the progression information (critical labels and progression days) with predicted risk scores. A larger C-index correlates with better progression prediction performance.

*A p-value from binomial test measures the difference in performance between the image and clinical combined model and other prediction models. A smaller p-value represents more significant difference between the combined model and other models.

**A p-value from log-rank test between high-risk and low-risk groups and Chi square values (χ^{2}) show a comparison of stratification performance of different models. A smaller p-value and larger χ^{2} correlate with better risk stratification performance.

References

1. 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):E197–206.

2. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012 Nov;30(9):1323–41.

3. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008 Sep;2(3):841–60.

4. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. arXiv:151204150 [cs] [Internet]. 2015 Dec 13 [cited 2020 Dec 11]; Available from: http://arxiv.org/abs/1512.04150

5. Fleming TR, Harrington DP. A class of hypothesis tests for one and two sample censored survival data. Communications in Statistics - Theory and Methods. 1981 Jan;10(8):763–94.