AI-based multi-modal integration (ScanCov scores) of clinical characteristics, lab tests and chest CTs improves COVID-19 outcome prediction of hospitalized patients

Nathalie Lassau[1,2], Samy Ammari[1,2], Emilie Chouzenoux[3], Hugo Gortais[4], Paul Herent[5], Matthieu Devilder[4], Samer Soliman[4], Olivier Meyrignac[4], Marie-Pauline Talabard[4], Jean-Philippe Lamarque[1,2], Remy Dubois[5], Nicolas Loiseau[5], Paul Trichelair[5], Etienne Bendjebbar[5], Gabriel Garcia[1], Corinne Balleyguier[1,2], Mansouria Merad[6], Annabelle Stoclin[7], Simon Jegou[5], Franck Griscelli[8], Nicolas Tetelboum[1], Yingping Li[2,3], Sagar Verma[3], Matthieu Terris[3], Tasnim Dardouri[3], Kavya Gupta[3], Ana Neacsu[3], Frank Chemouni[7], Meriem Sefta[5], Paul Jehanno[5], Imad Bousaid[9], Yannick Boursin[9], Emmanuel Planchet[9], Mikael Azoulay[9], Jocelyn Dachary[5], Fabien Brulport[5], Adrian Gonzalez[5], Olivier Dehaene[5], Jean-Baptiste Schiratti[5], Kathryn Schutte[5], Jean-Christophe Pesquet[3], Hugues Talbot[3], Elodie Pronier[5], Gilles Wainrib[5], Thomas Clozel[5], Fabrice Barlesi[6], Marie-France Bellin[2,4], Michael G. B. Blum[5*].

1.Imaging Department Gustave Roussy. Université Paris Saclay,Villejuif,F-94805

2.Biomaps. UMR1281 INSERM.CEA.CNRS.Université Paris-Saclay. Villejuif,F-94805

3.Centre de Vision Numérique, Université Paris-Saclay, CentraleSupélec, Inria, 91190 Gif-sur-Yvette, France

4.Radiology Department, Hôpital de Bicêtre – AP-HP, Université Paris Saclay, Le Kremlin-Bicêtre, France

5.Owkin Lab, Owkin, Inc. New York, NY USA

6.Département d'Oncologie Médicale, Gustave Roussy,Université Paris-Saclay, Villejuif, F-94805, France

7.Département de Soins Intensifs, Gustave Roussy, Université Paris-Saclay, Villejuif, F-94805, France

8.Département de Biologie, Gustave Roussy, Université Paris-Saclay, Villejuif, F-94805, France

9.Direction de la Transformation Numérique et des Systèmes d'Information, Gustave Roussy, 94800 Villejuif, France.

Corresponding author: michael.blum@owkin.com

The SARS-COV-2 pandemic has put pressure on Intensive Care Units, and made the identification of early predictors of disease severity a priority. We collected clinical, biological, chest CT scan data, and radiology reports from 1,003 coronavirus-infected patients from two French hospitals. Among 58 variables measured at admission, 11 clinical and 3 radiological variables were associated with severity. Next, using 506,341 chest CT images, we trained and evaluated deep learning models to segment the scans and reproduce radiologists' annotations. We also built CT image-based deep learning models that predicted severity better than models based on the radiologists' reports. Finally, we showed that adding CT scan information—either through radiologist lesion quantification or through deep learning—to clinical and biological data, improves prediction of severity. These findings show that CT scans contain novel and unique prognostic information, which we included in a 6-variable ScanCov severity score.

Introduction

Hospitalized COVID-19 patients are likely to develop severe outcomes requiring mechanical ventilation or high-flow oxygenation. Among hospitalized patients, 14 to 30% will require admission to an ICU, 12 to 33% will require mechanical ventilation, and 20% to 33% will die[1–4]. Detection at admission of patients at risk of severe outcomes is important to deliver proper care and to optimize use of limited intensive care unit (ICU) ressources[5].

Identification of hospitalized COVID-19 patients at risk for severe deterioration can be done using risk scores that combine several factors including age, sex, and comorbidities (CALL, COVID-GRAM)[6–11]. Some risk scores also include additional markers of severity such as the dyspnea symptom, clinical examination variables such as low oxygen saturation and elevated respiratory rate, as well as biological factors reflecting multi-organ failures such as elevated Lactate dehydrogenase (LDH) values[8,10,12–14].

Beyond clinical and biological variables, computerized tomography (CT) scans also contain prognostic information, as the degree of pulmonary inflammation is associated with clinical symptoms, and the amount of lung abnormality has been associated to severe evolution[15–19]. However, the extent to which CT scans at patient admission add prognostic information beyond what can be inferred from clinical and biological data is unresolved.

The objective of this study was to integrate clinical, biological and radiological data to predict the outcome of hospitalized patients. CT-scan information was included in multimodal scores either through deep learning models or using radiologist quantification of lesions.

Results

A total of 1,003 patients from Kremlin-Bicêtre (KB, Paris, France) and Gustave Roussy (IGR, Villejuif, France) were enrolled in the study. Clinical, biological, and CT scan images and reports were collected at hospital admission. There were 931 patients for which clinical, biological and CT-scan data were available (Supp Fig 1). A total of 506,341 images were analyzed for the 980 patients with CT-scans (average of 517 slices per scan). Radiologists annotated 17,873 images from 329 CT-scans. Summary statistics for the clinical, biological, and CT scan data are provided in Table 1.

	KB, N = 837			IGR, N = 150				
	Distribution	Association to severity p-value	Odds Ratio	Distribution	Association to severity p-value	Odds Ratio	Pooled p-value	Significant association
Outcomes								
Severe Outcome (Ventilation or O2 ≥ 15L or Death)	29% [817]	-		32% [144]	-		-	
Ventilation	12% [837]	-		11% [150]	-		-	
O2 ≥ 15L	15% [837]	-		11% [150]	-		-	
Death	17% [837]	-		17% [150]	-		-	
Clinical characteristics								
Age (years)	63 (52, 77) [833]	1.54E-09	1.66 (1.41,1.96)	61 (49, 71) [149]	1.29E-01	1.35 (0.92,1.98)	5.25E-10	*
Sex (males)	57% [837]	5.00E-05	1.95 (1.41,2.69)	49% [150]	8.67E-01	1.07 (0.51,2.23)	5.75E-05	*
Height (m)	1.7 (1.6, 1.8) [404]	1.21E-01	1.18 (0.96,1.45)	1.7 (1.6, 1.8) [109]	7.30E-01	1.08 (0.69,1.71)	1.12E-01	
BMI (kg/m^2)	27.0 (23.5, 31.1) [400]	1.63E-01	1.16 (0.94,1.42)	24.9 (21.5, 27.7) [108]	2.87E-01	1.28 (0.81,2)	1.19E-01	
Weight (kg)	75 (62, 90) [543]	4.35E-01	1.07 (0.9,1.28)	72 (60, 85) [119]	2.06E-01	1.31 (0.86,1.99)	3.23E-01	
Clinical examination								
Oxygen saturation (%)	95 (90, 97) [783]	2.44E-19	0.38 (0.31,0.47)	97.0 (94, 99) [132]	4.33E-04	0.35 (0.2,0.63)	2.91E-21	*
Diastolic pressure (mmHg)	80 (69, 90) [769]	3.07E-05	0.7 (0.59,0.83)	78 (69, 84) [138]	1.55E-01	0.75 (0.51,1.11)	1.35E-05	*
Respiratory rate (breaths/min)	25 (20, 30) [667]	8.85E-04	1.34 (1.13,1.59)	22 (18, 28) [67]	1.39E-02	3.37 (1.28,8.86)	2.14E-04	*
Systolic pressure (mmHg)	134 (118, 148) [769]	2.62E-02	0.83 (0.71,0.98)	124 (112, 137) [139]	1.37E-01	0.74 (0.5,1.1)	1.44E-02	
Cardiac frequency (bpm)	95 (82, 107) [771]	2.60E-01	0.91 (0.78,1.07)	90 (80, 103) [136]	2.33E-01	1.27 (0.86,1.87)	3.67E-01	
Body temperature (°C)	37.8 (37.0, 38.5) [788]	9.06E-01	0.99 (0.84,1.17)	37.3 (36.7, 38.0) [137]	5.44E-02	6.09 (0.97,38.36)	8.28E-01	
Symptoms								
Dyspnea	68% [837]	3.26E-03	1.69 (1.19,2.39)	57% [150]	1.88E-01	1.67 (0.78,3.59)	1.77E-03	
Chest pain	8.2% [837]	2.22E-03	0.29 (0.13,0.64)	5.3% [150]	7.68E-01	0.78 (0.15,4.04)	2.19E-03	
Confusion	9.4% [837]	1.70E-02	1.81 (1.11,2.95)	2.7% [150]	8.28E-02	7.62 (0.77,75.63)	8.03E-03	
Myalgia	24% [837]	8.53E-03	0.6 (0.41,0.88)	13% [150]	9.83E-01	0.99 (0.32,3.01)	9.49E-03	
Coughing	41% [837]	3.57E-02	0.71 (0.52,0.98)	31% [150]	7.64E-01	0.88 (0.4,1.97)	3.40E-02	
Dry quintuous cough	33% [837]	3.56E-01	0.86 (0.62,1.19)	20% [150]	6.62E-01	1.23 (0.48,3.17)	4.05E-01	
Headhache	5.3% [837]	5.23E-01	0.8 (0.4,1.6)	3.3% [150]	6.35E-01	0.58 (0.06,5.39)	4.76E-01	
Asthenia	17% [837]	3.49E-01	0.82 (0.54,1.24)	8.0% [150]	7.01E-02	3.18 (0.91,11.09)	5.43E-01	
Diarrhoea	8.0% [837]	5.91E-01	1.16 (0.67,2.03)	6.0% [150]	8.39E-01	1.2 (0.21,6.82)	5.72E-01	
Fever	51% [837]	7.57E-01	1.05 (0.77,1.42)	37% [150]	2.75E-01	1.53 (0.71,3.26)	6.21E-01	
Symptoms duration before examination	6 (3, 9) [771]	9.57E-01	1 (0.86,1.18)	4.0 (2.0, 7.8) [126]	4.76E-01	1.15 (0.78,1.68)	8.60E-01	
Comorbidities and smoking								
Chronic kidney disease	12% [837]	3.76E-05	2.51 (1.62,3.89)	7.3% [150]	1.06E-02	16.59 (1.93,142.84)	6.71E-06	*
Hypertension	45% [837]	1.05E-04	1.84 (1.35,2.51)	35% [150]	7.90E-01	1.11 (0.51,2.42)	1.11E-04	*
Asthma	8.7% [837]	6.31E-03	0.38 (0.19,0.76)	6.0% [150]	2.97E-01	0.32 (0.04,2.71)	4.09E-03	
Cardiac disease	22% [837]	2.80E-02	1.49 (1.04,2.13)	17% [150]	9.26E-01	1.05 (0.39,2.78)	2.92E-02	
Diabetes	24% [837]	5.68E-02	1.4 (0.99,1.98)	19% [150]	1.15E-01	2.1 (0.83,5.3)	3.16E-02	
Emphysema	6.0% [837]	1.64E-01	1.54 (0.84,2.83)	14% [150]	5.62E-03	4.12 (1.51,11.25)	6.40E-02	
Smoker	15% [837]	2.66E-01	1.27 (0.83,1.93)	25% [150]	1.17E-02	2.85 (1.26,6.44)	1.25E-01	
Chemotherapy	1.0% [837]	6.09E-01	1.46 (0.35,6.14)	39% [150]	9.94E-02	1.88 (0.89,3.98)	4.30E-01	
Dyslipidemia	16% [837]	6.51E-01	1.1 (0.73,1.66)	11% [150]	8.65E-02	2.67 (0.87,8.18)	4.58E-01	
Corticosteroids	3.9% [837]	3.02E-01	0.64 (0.27,1.5)	12% [150]	3.99E-02	3.14 (1.05,9.37)	5.10E-01	
NSAI	3.6% [837]	5.31E-01	1.28 (0.59,2.8)	0.7% [150]	9.88E-01	0 (0,Inf)	5.39E-01	
Cancer	7.3% [837]	4.85E-01	1.23 (0.69,2.18)	85% [150]	1.45E-01	0.47 (0.17,1.3)	6.63E-01	
Biological measures								
LDH (U/L)	341 (264, 453) [527]	9.55E-11	2.05 (1.65,2.54)	278.0 (203.8, 400.0) [134]	3.77E-03	2.36 (1.32,4.21)	6.05E-12	*
Urea (mM)	6.0 (4.2, 9.8) [739]	1.83E-09	1.7 (1.43,2.01)	5.1 (3.7, 6.9) [117]	1.28E-03	2.19 (1.36,3.52)	9.17E-11	*
CRP (mg/L)	69 (29, 130) [729]	1.97E-06	1.47 (1.25,1.72)	52.7 (15.1, 115.7) [142]	3.64E-02	1.48 (1.03,2.14)	4.48E-07	*
Neutrophils (10^9/L)	5.0 (3.5, 7.1) [746]	1.74E-04	1.36 (1.16,1.6)	4.4 (2.3, 7.1) [148]	4.51E-01	1.15 (0.8,1.64)	1.29E-04	*
Leucocytes (10^9/L)	6.7 (4.9, 9.0) [750]	1.37E-03	1.29 (1.1,1.51)	6.6 (4.1, 9.8) [144]	3.43E-01	1.19 (0.83,1.71)	9.11E-04	
Platelets (G/L)	206 (163, 260) [751]	7.71E-03	0.79 (0.66,0.94)	213.5 (156.0, 308.0) [144]	1.93E-01	0.76 (0.5,1.15)	4.37E-03	
Ferritin (µg/L)	627 (288, 1286) [296]	6.95E-02	1.26 (0.98,1.61)	567.0 (233.0, 1327.0) [125]	3.70E-02	1.52 (1.03,2.26)	3.16E-02	
Total bilirubin (µmol/L)	8 (6, 11) [691]	8.48E-02	1.15 (0.98,1.36)	9.0 (6.0, 13.0) [137]	3.25E-02	1.52 (1.04,2.23)	3.86E-02	
Conjugated bilirubin (µmol/L)	12 (9, 16) [57]	3.63E-01	1.32 (0.72,2.42)	2.0 (2.0, 5.0) [68]	2.29E-02	2 (1.1,3.65)	1.97E-01	
Lymphocytes (G/L)	1.0 (0.7, 1.4) [745]	2.96E-01	0.88 (0.69,1.12)	1.0 (0.6, 1.6) [148]	2.89E-01	0.8 (0.53,1.21)	2.25E-01	
Monocytes (10^9/L)	0.5 (0.3, 0.6) [746]	3.63E-01	0.93 (0.78,1.09)	0.5 (0.3, 0.8) [148]	1.64E-01	0.66 (0.36,1.19)	2.55E-01	
Haemoglobin (g/DL)	13.2 (12.1, 14.4) [751]	6.23E-01	1.04 (0.89,1.22)	11.1 (9.5, 12.8) [144]	3.79E-02	0.65 (0.43,0.98)	9.01E-01	
Creatine kinase (IU/L)	146 (84, 313) [638]	9.16E-01	0.99 (0.83,1.18)	70.5 (30.8, 157.2) [92]	2.36E-01	1.32 (0.83,2.1)	9.19E-01	
Radiological standardized report								
Disease extent 0/1/2/3/4/5	6.3/17/37/27/22/2.6% [806]	1.53E-19	2.37 (1.97,2.86)	14/36/24/7/10/9 [138]	1.21E-02	1.62 (1.11,2.37)	9.56E-21	*
Crazy paving	44% [799]	1.72E-08	2.5 (1.82,3.44)	40% [140]	2.72E-02	2.37 (1.1,5.11)	2.91E-09	*
Periphera topography	63% [749]	1.69E-04	0.54 (0.39,0.74)	32% [137]	1.77E-01	0.55 (0.23,1.31)	8.21E-05	*
Inferior predominance	40% [732]	3.35E-02	1.42 (1.03,1.97)	43% [136]	7.81E-01	0.9 (0.42,1.93)	4.08E-02	
Ground glass opacities	91% [824]	1.10E-01	1.64 (0.89,3.02)	76% [140]	3.66E-01	1.55 (0.6,3.98)	8.32E-02	
Consolidation	69% [773]	1.33E-01	1.32 (0.92,1.89)	39% [140]	7.56E-01	1.13 (0.53,2.41)	1.25E-01	
Ground glass opacities rounded	15% [750]	1.07E-01	0.68 (0.42,1.09)	7.5% [107]	1.56E-01	3.1 (0.65,14.75)	1.80E-01	
Consolidation rounded	17% [535]	5.67E-01	1.16 (0.7,1.9)	18% [55]	9.11E-01	0.92 (0.21,4.09)	5.86E-01	
Other radiological patterns								
Cardiomegaly	25% [115]	4.16E-02	2.48 (1.03,5.92)	24% [76]	6.84E-01	0.77 (0.22,2.72)	5.29E-02	
Splenomegaly	2.6% [115]	2.41E-01	4.29 (0.38,48.86)	9.2% [76]	1.19E-02	17.33 (1.88,160)	1.12E-01	
Hepatomegaly	8.7% [115]	6.10E-01	1.41 (0.37,5.35)	17% [76]	1.33E-02	5.17 (1.41,18.99)	3.51E-01	
COVID-specific treatment								
Corticosteroids	1.0% [837]	-	-	4.7% [150]	-	-	-	
Hydroxychloroquine	4.1% [837]	-	-	34% [150]	-	-	-	
Other immunomodulator	2.5% [837]	-	-	0% [144]	-	-	-	
Lopinavir-Ritonavir	1.4% [837]	-	-	2.0% [150]	-	-	-	
Anti IL6	3.2% [837]	-	-	4.7% [150]	-	-	-	
Interferon	0.6% [837]	-	-	0% [144]	-	-	-	

Table 1: Population description for the KB and IGR hospitals and association between variables measured at admission and severity. Among the 1,003 patients of the study, biological and clinical variables were available for 987 individuals. Categorical variables are expressed as percentages [available]. Continuous variables are shown as median (IQR) [available]. Association with severity are reported with p-values for each center and p-value were combined with Stouffer's method. A star (*) in the column entitled "Significant association" indicates that the variable is significantly associated with severity after Bonferroni adjustment to account for multiple testing across 58 variables (treatments are excluded). For continuous variables, odds ratios are computed for an increase of one standard deviation of the continuous variable.

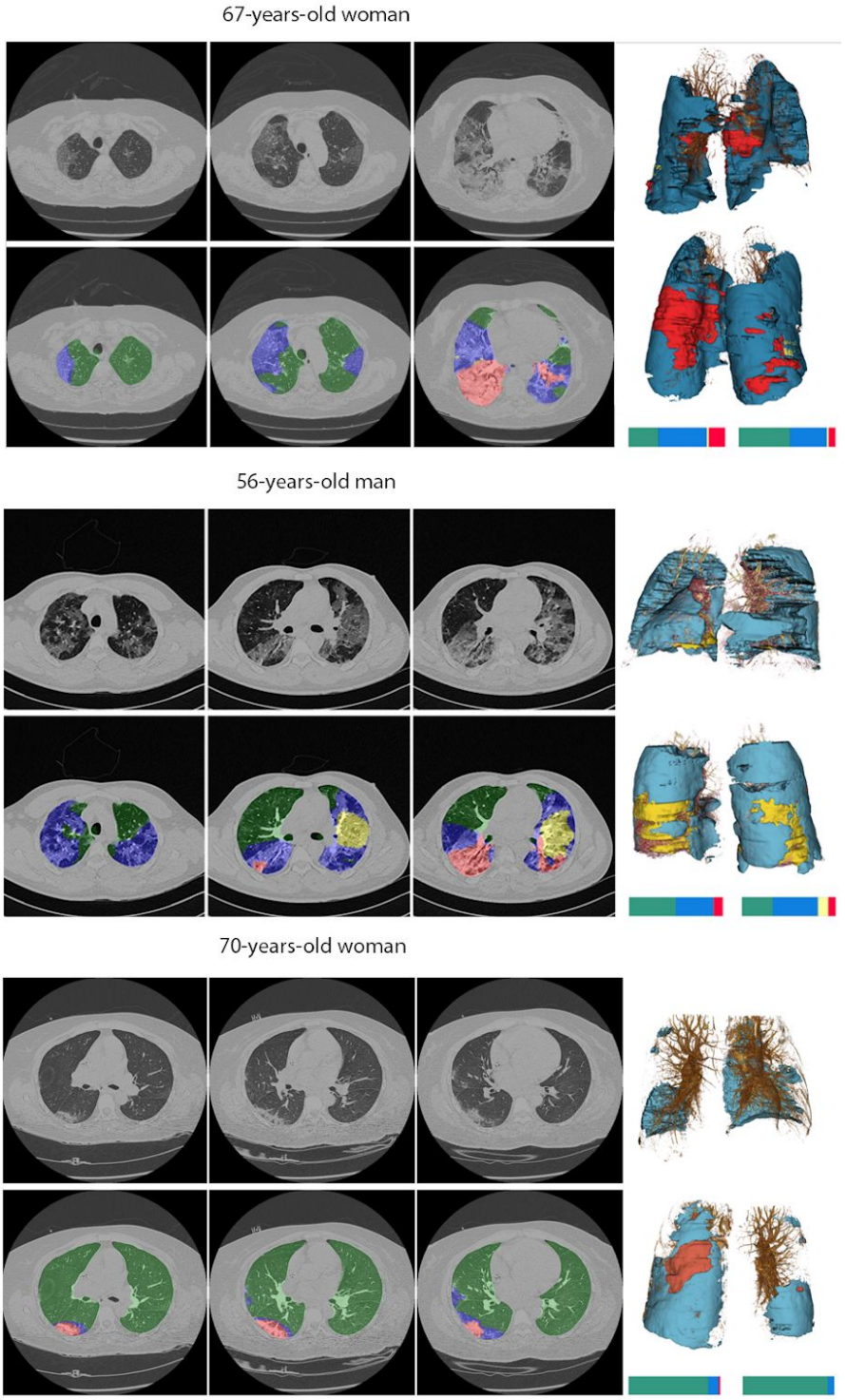

Figure 1: **Axial chest CT scans and segmentation results** COVID-19 radiology patterns, as provided by the model *AI-segment*, for 3 patients with COVID-19. Green/transparent: sane lung; blue: GGO; yellow : crazy paving; red: consolidation. (Top) 67-year-old woman with diffuse distribution, and multiple large regions of subpleural GGO with consolidation to the right and left lower lobe. Estimated disease extent by AI: 69%/47% (right/left). Radiologist report: critical stage of COVID-19 (stage 5). (Middle) 56-year-old man, with diffuse distribution and multiple large regions of subpleural GGO with superimposed intralobular and interlobular septal thickening (crazy paving). Estimated disease extent by AI: 51%/68% (right/left). Radiologist report: severe stage of COVID-19 (stage 4). (Bottom) 70-year-old woman, with minimal impairment, and multiple small regions of subpleural GGO with consolidation to the right lower lobe. Estimated disease extent 13%/7% (left/right). Radiologist report: moderate stage of COVID-19 (stage 2).

Variables associated with severity

We first evaluated how clinical and biological variables measured at admission were associated with future severe progression, which we defined as an oxygen flow rate of 15 L/min or higher and/or the need for mechanical ventilation and/or patient death[20]. This definition of severe progression corresponds to a score of 5 or more according to the World Health Organization evaluation of severity on a 1 to 10 scale. We computed the severity odds ratios for each individual variable, and at each hospital center (Table 1 and Supp Fig 2). When combining association results from the two centers, we found 11 variables significantly associated with severity (P <0.05/58 to account for testing 58 variables, Table 1 and Supp Fig 2): **age** (Odds Ratio [OR] KB 1.66 (1.41-1.96), OR IGR 1.35 (0.92-1.98), $P_{Stouffer}$ = 5.25e-10), **sex** (OR KB 1.95 (1.41-2.69), OR IGR 1.07 (0.51-2.23), $P_{Stouffer}$ = 5.75e-05), **hypertension** (OR KB 1.84 (1.35-2.51), OR IGR 1.11 (0.51-2.42), $P_{Stouffer}$ = 1.11e-04), **chronic kidney disease** (OR KB 2.51 (1.62-3.89), OR IGR 16.59 (1.93-142.84), $P_{Stouffer}$ = 6.71e-06), **respiratory rate** (OR KB 1.34 (1.13-1.59), OR IGR 3.37 (1.28-8.86), $P_{Stouffer}$ = 2.14e-04), **oxygen saturation** (OR KB 0.38 (0.31-0.47), OR IGR 0.35 (0.20-0.63), $P_{Stouffer}$ = 2.91e-21), **diastolic pressure** (OR KB 0.70 (0.59-0.83), OR IGR 0.75 (0.51-1.11), $P_{Stouffer}$ = 1.35e-05), **CRP** (OR KB 1.47 (1.25-1.72), OR IGR 1.48 (1.03-2.14), $P_{Stouffer}$ = 4.48e-07), **LDH** (OR KB 2.05 (1.65-2.54), OR IGR 2.36 (1.32-4.21), $P_{Stouffer}$ = 6.05e-12), **polynuclear neutrophil** (OR KB 1.36 (1.13-1.60), OR IGR 1.15 (0.80-1.64), $P_{Stouffer}$ = 1.29e-04), and **urea** (OR KB 1.70 (1.43-2.01), OR IGR 2.19 (1.36-3.52), $P_{Stouffer}$ = 9.17e-11).

We then assessed the predictive value of features from admission radiology reports, and found three significant features: (i) **extent of disease** (OR KB 2.37 (1.97-2.86), OR IGR 1.62 (1.11-2.37), $P_{Stouffer}$ = 9.56e-21) and (ii) **crazy paving** *(*OR KB 2.50 (1.82-3.44), OR IGR 2.37 (1.10-5.11), $P_{Stouffer}$ = 3.10e-09), associated with greater severity, and (iii) **peripheral topography**, associated with lesser severity (OR KB 0.54 (0.39-0.74), OR IGR 0.55 (0.23-1.31), $P_{Stouffer}$ = 8.21e-05).

Segmentation of CT-scans

We next trained the deep neural network *AI-segment* (Supp Fig 3) to segment radiological patterns and provide automatic quantification[21,22] of their volume, expressed as a percentage of the full lung volume. These patterns included the three distinguishable features that appear as disease severity progresses: ground glass opacity (GGO), crazy paving, and finally consolidation. *AI-segment* was trained on 184 patients from KB hospital (8 fully annotated scans, 176 partially annotated ones) and evaluated on 145 patients from IGR hospital (14 fully annotated scans and 131 partially annotated ones). To evaluate *AI-segment*, we first compared its performance to that of radiologists manual annotation. *AI-segment* discriminated lung regions from regions outside of the lung with an accuracy of 99.9% when evaluated on the fully annotated scans. Within the lung, the model's ability to discriminate between lesions and healthy areas had F1 values of 0.85 and 0.98 on partially and fully annotated scans. In the fully annotated scans, the predicted volumes of each lesion type had relative errors (median [min-max]) of 3.77% [0.054%-14%] for GGO, 0.96% [0.058%-4.4%] for consolidation, and 5.92% [0.41%-13%] for sane lung (no crazy paving

was present in these scans). We next compared *AI-segment* to the information contained in the radiology reports. The F1 score measuring the ability of *AI-segment* to detect the presence of a lesion type per patient, was of 0.88 for GGO, 0.65 for crazy paving, and 0.75 for consolidation (Supp Table 1). Correlation between quantification of the proportion of lesions with *AI-segment* and the radiologist evaluation was of 0.56 (Supp Fig 5). *AI-segment* visual results were also consistent with radiologist observations (see Figure 1 for three representative cases). We lastly evaluated to what extent *AI-segment* provided biomarkers of future severity. We found that severity was significantly associated to GGO extent (OR KB 0.64 (0.54,0.76), 0.77 (0.54,1.10), $P_{Stouffer}$ = 1.94e-07), crazy paving extent (OR KB 1.47 (1.20-1.79), OR IGR 1.31 (0.92,1.87), $P_{Stouffer}$ = 6.70e-05), consolidation extent(OR KB 1.46 (1.23,1.73), 1.27 (0.89,1.82), $P_{Stouffer}$ = 7.61e-06) as well as total disease extent (OR KB 2.11 (1.74,2.55), OR IGR 1.90 (1.30,2.79), $P_{Stouffer}$ = 7.66e-16)(accounting for multiple testing). These correlations were observed in the larger KB dataset, but were not found in the IGR dataset (Supp Table 2).

Prognostic models based on CT-scan only

We next evaluated the prognostic value of variables extracted from CT scans through three different models. The first model called *report* combined variables from the radiological report using logistic regression. The second was based on the lesion volumes computed by *AI-segment* and variables were again combined with logistic regression. The third called *AI-severity* used a weakly supervised approach with no radiologist-provided annotations (Supp Fig 4)[23]. All three models were trained on 646 KB patients, validated on 150 KB patients and on the independent IGR dataset of 135 patients (Figure 2). On the validation set from KB hospital, *AI-severity* outperformed *report* ($AUC_{AI\text{-}severity}$ = 0.76 (0.66,0.85), $AUC_{AI\text{-}segment}$ = 0.67(0.56,0.77), AUC_{report} = 0.71 (0.62,0.80)). On the independent IGR validation set, both *AI-segment* and *AI-severity* outperformed the model *report* ($AUC_{AI\text{-}severity}$ = 0.75 (0.65,0.84), $AUC_{AI\text{-}segment}$ = 0.70 (0.59,0.80), AUC_{report} = 0.65 (0.54,0.75)). When considering alternatives outcomes consisting of either death, or death or admission to ICU, *AI-severity* and *AI-segment* were also superior to *report* in terms of AUC (Supp Table 3).

To interpret the weakly supervised *AI-severity* model, and understand what it detects within the CT scans, we evaluated to what extent the features extracted by *AI-severity* (internal representation) could predict clinical and radiological variables. To this end, we trained a new logistic regression with *AI-severity's* extracted features as input, and some clinical and radiological variables as output. AUC on the KB validation set was 0.93 (C.I. = (0.88,0.97)) for disease extent (threshold >2), 0.78 (C.I. = (0.70,0.85)) for crazy paving, 0.64 for condensation (C.I. = (0.53,0.74)) and 0.80 for GGO (C.I. = (0.65,0.94)) (Supp Table 4). It was also possible to relate internal representations of the neural networks to clinical variables. We obtained an AUC of 0.88 (C.I. = (0.82,0.94)) for predicting an age strictly larger than 60 year-old, an AUC of 0.93 (C.I. = (0.89,0.97)) for sex, and of 0.76 (C.I. = (0.68,0.84)) for predicting an oxygen saturation larger than 90%. As a comparison, a logistic regression trained on the variables from the radiology report obtained only AUC scores of 0.70 (C.I. = (0.61,0.78)) for age, 0.57 (C.I. = (0.48, 0.67)) for sex and of 0.68 (C.I. = (0.58, 0.77)) for oxygen saturation. Simply put, this analysis shows that the internal representation

of the *AI-severity* neural network captures clinical features from the lung CTs, such as sex or age, on top of the known COVID-19 radiology features.

***Multimodal prognostic models and ScanCov* score**

Lastly, we evaluated whether CT scans have a prognostic value beyond what can be inferred from clinical and biological characteristics alone. To this end, we sought to compare the performance of trimodal CT scan/clinical/biological models to a bimodal clinical/biological model (*C & B*). Using a greedy search approach to include optimal variables, we therefore incorporated clinical and biological variables into *report* and named the resulting trimodal model the *ScanCov* score. Coefficients and transformations required to compute the 6-variable *ScanCov* score are available in Supp Table 6. Through the same method, we also made a trimodal version of *AI-segment*, and *AI-severity* (Supp Fig 6, Supp Table 5). We evaluated the models' performances on three outcomes: the initial WHO-defined high severity outcome of "oxygen flow rate of 15 L/min or higher, or need for mechanical ventilation, or death", as well as two other outcomes "death or ICU admission", and "death". For each outcome and validation set, both *ScanCov* and *AI-severity* performed better than the bimodal biological/clinical *C & B* model (Figures 2 & 3, Supp Table 3). The gain of performance when compared to the *C & B* model was larger for the KB hospital *(*median AUC increase of 4.0% for *AI-severity* and of 3.6% for *ScanCov)* than for the IGR hospital (median AUC increase of 1.5% for *AI-severity* and of 0.4% for *ScanCov)*. For the model *AI-segment*, the median increase of AUC was of 0.5% for the KB hospital and of 1.9% for the IGR hospital.

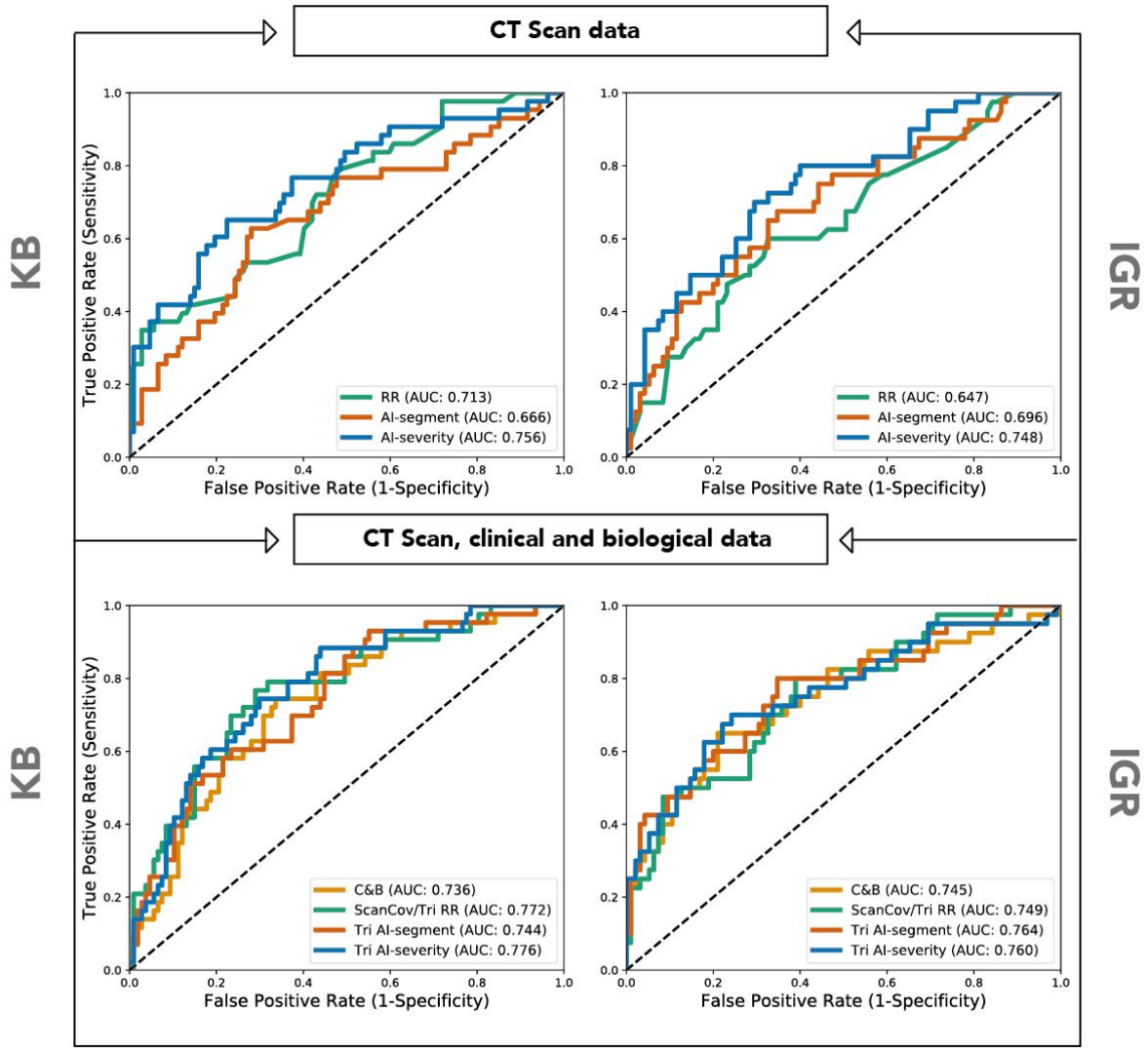

Figure 2: Receiver operating characteristic (ROC) curves for the models that predict severity. Models were evaluated on two distinct validation sets consisting of 150 patients from KB (left panels) and 135 patients from IGR (right panels). The model *RR* denotes the model based on the variables of the radiologist report, the model *C & B* denotes the model based on clinical and biological data only, and the prefix Tri stands for Trimodal when clinical and biological data were included in addition to CT-scan data. *AI-segment* and *AI-severity* denote two models based on deep learning variables extracted from CT-scans.

The *ScanCov* and *AI-severity* models also outperformed other previously published severity or mortality scores (Figure 3, Supp Fig 7, Supp Table 3). The median difference (averaged over outcomes) between the AUC of *AI-severity* and of other scores ranged between 5% (COVID-GRAM) and 15% (CALL) at KB and between 10% (COVID-GRAM) and 26%[15] at IGR. The median difference (averaged over outcomes) between the AUC of *AI-segment* and of other scores ranged between 2% (COVID-GRAM) and 12% (CALL) at KB and between 5% (COVID-GRAM) and 24%[15] at IGR. Similarly, the median difference (averaged over outcomes) between the AUC of *ScanCov* and of other scores ranged between 4% (COVID-GRAM) and 14% (CALL) at KB and between 5% (COVID-GRAM) and 24%[15] at IGR. Among alternative scores, the COVID-GRAM score provided the largest value of

median AUC (Figure 3, Supp Table 3). When averaging AUC (median) over outcomes at KB, the range of AUC increase when comparing COVID-GRAM to other scores was between 0.5% (MIT analytics) and 10% (CALL); range of AUC increase at IGR was between 7% (MIT analytics) and 10%[15]. The COVID-GRAM score was also the only alternative scoring system we considered that includes CT scan information.

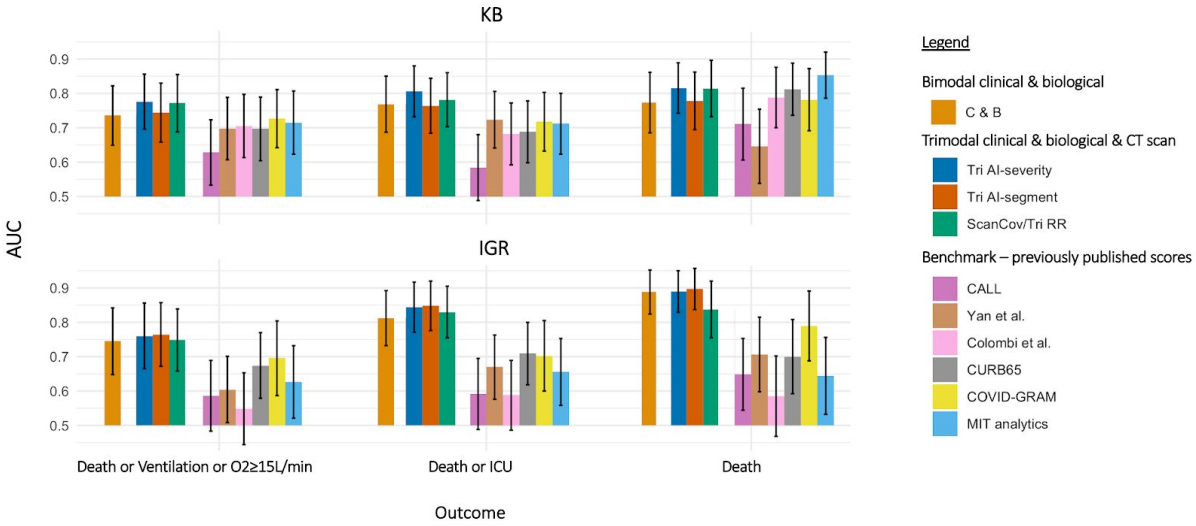

Figure 3: **AUC values of the different models for the three distinct outcomes.** The *C & B* (bimodal model with clinical and biological information), *Tri AI-severity, Tri AI-segment* (trimodal models with deep learning to extract information from CT scans) and *ScanCov* 6-variable trimodal model were trained on 646 patients from KB. These models were trained using the severity outcome defined as an oxygen flow rate of 15 L/min or higher, the need for mechanical ventilation, or death. When evaluating these 4 models on the alternative outcomes, models were not trained again. Results are reported on the validation set from KB (150 patients) and the external validation set from IGR (135 patients). Error bars represent the 95% confidence intervals.

Of all the features within the radiologists' report, disease extent was the most strongly associated to prognosis (Table 1). We therefore further investigated this feature to confirm that it brings additional prognostic information that is not otherwise captured in any clinical or biological variable. In the KB dataset, the three variables that were the most correlated with disease extent were LDH (r = 0.52, C.I. = (0.45,0.58)), CRP (r = 0.45, C.I. = (0.39,0.51)), and oxygen saturation (r = -0.43, C.I. = (-0.49,-0.37)) (Supp Table 7). We then regressed the severity outcome with disease extent and the three correlated variables and found that significant predictors included oxygen saturation (P = 1.57e-07) and disease extent (P = 0.01), whereas statistical evidence for association was weak for LDH (P = 0.06) and absent for CRP (P = 0.26). The statistical evidence for association between disease extent and severity was also found (P = 9.85e-08) when accounting for the five additional variables of the *ScanCov* score, which were also significantly related with the outcome (Age P = 1.49e-06, Oxygen saturation P = 2.83e-08, Sex P = 0.035, Platelet P = 0.001, Urea = 9.77e-05). This confirms that the radiological feature of disease extent brings unique prognostic information.

To further evaluate the *ScanCov* score, individuals in the top tercile were assigned in a high risk group. We found that the survival function of the individuals at high risk was significantly different from the survival function of the other individuals (Figure 4, P = 2.90e-07 at KB, P =

5.38e-08 at IGR for a log-rank test). When considering a binary classification consisting of a high-risk group and a medium or low risk group, we obtained positive predictive values (or precision) of 54% (KB) and 68% (IGR), negative predictive values of 85% (KB) and 80% (IGR), specificities of 78% (KB) and 91% (IGR), and sensitivities of 65% (KB) and 48% (IGR).

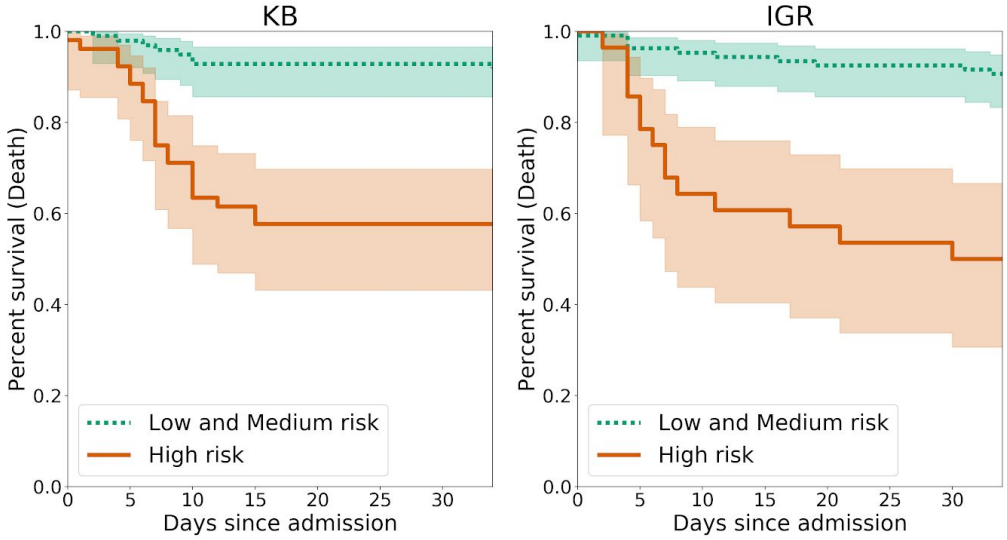

Figure 4: KM curves for the high risk individuals and the ones with low or medium risk according to the *ScanCov* score. The threshold to assign individuals into a high risk group was the ⅔-quantile of the *ScanCov* scores computed for patients of the KB training set. Kaplan-Meier curves were obtained for individuals of the KB validation set (left panel) and IGR validation set (right panel). P-values for the log-rank test were equal to 2.90e-07 (KB) and 5.38e-08 (IGR). The two terciles used to determine threshold values for low, medium and high risk groups were equal to 0.144, and 0.369.

Discussion

Taken together, these results show that unique future disease severity markers are present within routine CT scans performed at admission, and that these scans provide useful and interpretable elements for prognosis.

Looking back on the prognostic clinical and biological variables, we found 11 of these significantly associated with severe evolution, which is consistent with previous studies[14,30,37]. First, looking at clinical characteristics, we confirmed that male and older persons are more at risk[24]. Although BMI is a known risk clinical factor for severe COVID[24], it was not associated with severity here. Discussion with clinicians however indicated that the data capture may have been biased, with emergency room doctors inputting height and weight more frequently for obese patients. Second, looking at clinical examination variables, we found that respiratory rate, diastolic pressure, and oxygen saturation are clinical variables associated with severity. These associations may reflect physician decisions taken for ICU triage. Inclusion criteria for critical care triage include (i) requirement for invasive ventilatory

support characterized by an oxygen saturation lower than 90%, or by respiratory failure, or (ii) requirement for vasopressors characterized by hypotension and low blood pressure[33]. Third, looking at comorbidities, we confirmed the results of several meta-analyses[28,30–32] that showed that chronic kidney disease and hypertension are linked to severity. We however did not find significant associations for other comorbidities previous associated with severity, such as diabetes, and cardiovascular diseases[28,29]. While we expected cancer patients to have more severe outcomes because they are generally older, with multiple co-morbidities and often in a treatment-induced immunosuppressive state[25–27], we did not find this association. Several factors can explain this. Each cohort was not optimally balanced to conclusively study the association between cancer and severity: IGR admitted mostly cancer patients (80% of the patients) while KB admitted very few cancer patients (7%). Fourth, looking at COVID symptoms, we did not find any significantly associated with severity. Dyspnea is a prominent symptom that has been repeatedly associated with severity and our results are compatible with a positive association with severity but we may lack a large-enough sample size to be significant[6,34,35]. Last, looking at biological measures, we found that inflammatory biomarkers, LDH and CRP are related to severity[13,36,37]. We also found association with neutrophil and urea, the later being explained by the fact that high urea is indicative of kidney disfunction. Thrombocytopenia (low platelet count) was not significantly associated to severity, possibly because of lack of statistical power and stringent correction for multiple testing, but association betwen thrombocytopenia and severity was in the expected direction and platelet counts are included in the 6-variable *ScanCov* score and in the trimodal models *AI-segment* and *AI-severity* .

Beyond these clinical and biological variables, chest CT-scans provided additional markers of disease severity. Significant features include the total extent of lesions, the presence of crazy paving pattern lesions, and the proportion of consolidation lesions when measured with automatic segmentation. Although the extent of disease severity and consolidation are known to be associated with severity[15,18,38–43], our study discovered its association with crazy paving, a precursor of consolidation lesions. Initial damages to the alveoli, as well as protein and fibrous exudation, explain the early onset of GGO. As the disease progresses, more and more inflammatory cells infiltrate the alveoli and interstitial space, followed by diffuse alveolar lesions and the formation of a hyaline membrane, which results in a crazy paving appearance, which is then followed by consolidation on the CT examination[44,45]. Correlation results between the proportion of each lesion type and severity reflects this sequencing, as GGO proportions are negatively related with severity, while crazy paving and consolidation proportions are positively correlated to severity (Supp Table 2).

Compared to a radiologist's reporting and quantification of lesions, there are several advantages to capturing CT-scan information through deep learning models. Good reproducibility is a key element for imaging biomarkers such as disease extent, and visual inspection of images introduces variability that can hinder its clinical application[46]. Additionally, the consolidation feature, which has been repeatedly associated with COVID severity[34,38,41–43], was not found to be associated with severity with a simple presence/absence radiologist coding, whereas correlation was evident once pulmonary consolidation was quantified with automatic segmentation. Another advantage is that radiologists are faced with the challenge that large numbers of cases must be read, annotated and prioritized in a COVID-19 pandemic. AI analysis of radiological images has

the potential to reduce this burden and speed up their reading time. Finally, unimodal prognosis scores obtained with deep learning models trained on CT-scans are more predictive of severity than manually extracted radiological features. We indeed showed that internal representation of the *AI-severity* neural network captures clinical information from CT scans, and this can can be particularly useful when some clinical or lab measurements are missing.

Our reported prognostic values for CT-scan-based models (AUC range of 0.70 - 0.80) are lower than the 0.85 AUC reported in the previously published Zhang et al study[16]. We hypothesize that this is due to use of different outcome definitions, as well as different patient characteristics in the study cohorts (age, severity at admission, etc). Hospital admission criteria vary between countries and hospitals; for instance, proportion of deaths in our French KB and IGR cohorts was of 16-17%, while it was of 39% in the Zhang et al study[16]. When applying other previously published scores to the KB and IGR datasets, we found smaller AUC scores than reported values in the original papers. This difference can again be explained by differing patient characteristics, and different measures of severity between studies[7,10,15,6,9]

Our evaluation of the different trimodal models that included CT scan information in addition to clinical and biological information revealed the added prognostic value of CT scans. Interestingly, while CT scan disease extent was correlated to biological and clinical severity biomarkers such as CRP levels, tissue damage (LDH) and oxygenation —highlighting some information redundancy between data modalities[34,47–4]—disease extent was still significantly associated with severity even after accounting for these other severity markers, confirming the unique added value of CT-scans. Beyond AI modeling, our study shows that the 6-variable *ScanCov* score integrating a radiological quantification of lesions with key clinical and biological variables provides accurate severity predictions, and can rapidly become a reference patient scoring approach.

Methods

Description of the retrospective study

Data including CT-scans, were collected at two French hospitals (Kremlin Bicêtre Hospital, APHP, Paris denotes as KB and Gustave Roussy Hospital, Villejuif denoted as IGR). CT scans, clinical, and biological data were collected in the first 2 days after hospital admission. This study has received approval of ethic committees from the two hospitals and authors submitted a declaration to the National Commission of Data Processing and Liberties (N° INDS MR5413020420, CNIL) in order to get registered in the medical studies database and respect the General Regulation on Data Protection (RGPD) requirements. An information letter was sent to all patients included in the study. We stopped to update information about patient status on the 5th of May. Among the 1,003 patients of the study, two patients asked to be excluded from the study.

Inclusion criteria were (1) date of admission at hospital (from the 12th of February to the 20th of March at Kremin Bicêtre and from the 2nd of March to the 24th of April at Institut Gustave

Roussy) and (2) a positive diagnosis of COVID-19. Patients were considered positive either because of a positive RT-PCR (real-time fluorescence polymerase chain reaction) based on nasal or lower respiratory tract specimens or a CT scan with a typical appearance of COVID-19 as defined by the ACR criteria for negative RT-PCR patients[50]. Children and pregnant women were excluded from the study.

The clinical and laboratory data were obtained from detailed medical records, cleaned and formatted retrospectively by 10 radiologists with 3 to 20 years of experience (5 radiologists at GR and 5 at KB). Data include demographic variables: age and sex, variables from the clinical examination include: body weight and height, body mass index, heart rate, body temperature, oxygen saturation, blood pressure, respiratory rate, and a list of symptoms including cough, sputum, chest pain, muscle pain, abdominal pain or diarrhoea, and dyspnea. Health and medical history data include presence or absence of comorbidities (systemic hypertension, diabetes mellitus, asthma, heart disease, emphysema, immunodeficiency) and smoker status. Laboratory data include conjugated alanine, bilirubin, total bilirubin, creatine kinase, CRP, ferritin, haemoglobin, LDH, leucocytes, lymphocyte, monocyte, platelet, polynuclear neutrophil, and urea.

CT scan acquisition

CT scan data were available for 980 patients representing a total of 506,341 images (517 slices per patient on average). Summary statistics for the clinical, biological, and CT scan data are provided in Table 1. Three different models of CT scanners were used : two General Electric CT scanners (Discovery CT750 HD and Optima 660 GE Medical Systems, Milwaukee, USA) and a Siemens CT scanner (Somatom Drive; Siemens Medical Solutions, Forchheim). All patients were scanned in a supine position during breath-holding at full inspiration. The acquisition and reconstruction parameters were of 120kV tube voltage with automatic tube current modulation (100-350 mAs), 1 mm slice thickness without interslice gap, using filtered-back-projection (FBP) reconstruction (SOMATOM Drive) or blended FBP/iterative reconstruction (Discovery or Optima). Axial images with slice thickness of 1 mm were used for coronal and sagittal reconstructions.

Radiology reports

COVID-19 associated CT imaging features were obtained from radiologist reports that follow the guidelines of several scientific societies of radiology (French SFR, STR, ACR, RSNA) regarding the reporting of chest CT findings related to COVID-19 [50]. The template of the radiologist report (https://ebulletin.radiologie.fr/actualites-covid-19/compte-rendu-tdm-thoracique-iv-0) was available the 17th of March and the reports were completed retrospectively for the patients who were admitted to the hospital before that date. CT imaging characteristics were evaluated to provide the five following variables : (i) ground glass opacity (GGO) (rounded / non rounded / absent) that is defined as an increase in lung density not sufficient to obscure vessels or preservation of bronchial and vascular margins, (ii) consolidation (rounded / non rounded / absent) that occurs when parenchymal opacification is dense enough to obscure

the vessels' margins and airway walls and other parenchymal structures, (iii) the crazy-paving pattern (present/absent) that is defined as ground-glass opacification with associated interlobular septal thickening[51], (iv) peripheral topography (present/absent) that corresponds to the spatial distribution of lesions in the one-third external part of the lung, and (v) inferior predominance (present/absent) that is defined as a predominance of lesions located in the lower segments of the lung. A rounded pattern (for GGO and consolidation) is defined as a lesion presenting a well delineated shape. In addition to the five CT imaging features, radiologists assessed the extent of lung lesions according to the evaluation criteria established by the French Society of Radiology (SFR)[52]. Disease extent can be: absent / minimal (<10%) / moderate (10-25%) / extensive (25-50%) / severe (>50%) / critical >75%. The coding absent / minimal / moderate extensive / severe / critical was based on a quantitative variable with values of 0 / 1 / 2 / 3 / 4 / 5. Variables were automatically extracted from the report using optical character recognition.

Annotation scenario of CT scans by radiologists

Two radiologists (4 and 9 years of experience) examined and annotated 307 anonymized chest scans independently and without access to the patient's clinic or COVID-19 PCR results. All CT images were viewed with lung window parameters (width, 1500 HU; level, -550 HU) using the SPYD software developed by Owkin. Regions of interest were annotated by the radiologists in four distinct classes : healthy pulmonary parenchyma, ground glass opacity, consolidation, crazy-paving. The presence of organomegaly was also notified when present, as a binary class. When multiple CT images were available for a single patient, the image to analyze was selected using the SPYD software. One AI and imaging PhD student also provided full 3D annotation of the four classes on 22 anonymized chest scans using the 3D Slicer software.

Statistical analysis

When detecting association with the severity outcome, odds ratio and P-values (two-sided tests) were computed separately for each hospital using logistic regression (*glm* function of the R statistical software). P-values from the two different hospitals were pooled using the Stouffer meta-analysis formula accounting for the two different sample sizes. For association between severity and each variable, we considered Bonferroni correction accounting for 58 variables and 62 variables when also considering imaging markers obtained with *AI-segment*. To compute confidence intervals for AUC values, we considered DeLong method [53]. Survival functions were obtained using Kaplan-Meier estimators.

The AI-*segment* pipeline for lesion segmentation from CT scans was based on 3 segmentation networks: 3D Resnet50[54] , 2.5D U-Net, and 2D U-Net [55]. U-Net consists of convolution, max pooling, ReLU activations, concatenation and up-sampling layers with sections: contraction, bottleneck, and expansion. ResNet contains convolutions, max pooling, batch normalization, and ReLU layers that are grouped in multiple bottleneck blocks. All models were trained on CT scans provided by Kremlin-Bicêtre (KB) and evaluated on annotated CT scans from Institut Gustave Roussy (IGR). The dataset was divided into two categories: Fully Annotated Scans (FAS) composed of 22 scans (8 from KB

and 14 from IGR) and Partially Annotated Scans (PAS) composed of 307 scans (176 from KB and 131 from IGR). PAS contains a total of 7,374 annotated slices and 24,476,521 annotated pixels, i.e. 24 slices per PAS and 3,319 pixels annotated per slice on average.

2D U-Net was trained for left/right lung segmentation and 3D ResNet and 2.5D U-Net were used for lesion segmentation. 3D ResNet50 was trained on 8 KB FAS (i.e. 3,704 slices). Inputs for the 3D ResNet consist of a height and a width of 128, and a depth of 32. We initialized the 3D ResNet with pretrained weights[56]. We then trained the network with Stochastic Gradient Descent for parameter optimization and an initial learning rate of 0.1 with a decay factor of 0.1 every 20 epochs. The network was trained for a total of 100 epochs. For the 2.5D U-Net, we first pretrained the network on a left-right lung segmentation task using the LCTCS dataset[57]. The network was then trained on the KB dataset using Adam optimization algorithm with a learning rate, weight decay, gradient clipping and learning rate decay parameters of 1e-3, 1e-8, 1e-1, and 0.1 (applied at epochs 90 and 150) for 300 epochs. While the validation set remains the same as when evaluating the 3D resnet50 model, 176 KB PAS scans were added to the 8 KB FAS, in the training set. PAS were only added to the 2.5D U-Net training set due to the incompleteness of the annotated volume in the scans which would not satisfy the volumetric requirements of the 3D ResNet50 input. Finally, for the left/right lung segmentation, the 2D U-Net was trained on the 8 KB FAS. Similarly to 2.5D U-Net, Adam optimization algorithm was used with a learning rate, weight decay, gradient clipping, learning rate decay, and number of epochs of 1e-3, 1e-8, 1e-1, 0.1 (applied at epoch 70), and 104. Both 2.5D U-Net and 2D U-Net used affine transformation and contrast change for data augmentation while 3D ResNet50 used affine transformation, contrast change, thin plate splines, and flipping. 3D ResNet and 2.5D U-Net are trained through the minimization of a cross entropy loss and 2D U-Net minimized a binary cross entropy loss. All training was performed on NVIDIA Tesla V100 GPUs using Pytorch as a coding framework. During the validation phase, ensemble inference[58] was performed on all available scans by averaging lesion masks, which were predicted from the 3D ResNet and 2.5D U-Net models, using arithmetic mean.

We evaluated *AI-segment* on three distinct aspects. First, we evaluated its ability to perform accurate segmentation. To this aim, we computed F1 scores for the PAS (partially annotated scans) and FAS (fully annotated scans), of the IGR test set, when discriminating lesions versus sane areas inside the lung. Micro-averaging was used to limit the effect of class imbalance for the three different lesion types. We also reported the accuracy to discriminate background versus lung regions using FAS where background regions outside of the lung were annotated. Second, we evaluated its ability to estimate the proportion of each lesion type per scan. To this aim, we computed the median, minimum and maximum of the absolute value of the difference between the ground truth percentage of each lesion type obtained from radiologists' annotations and the estimated ones, on the 14 available FAS of the IGR dataset. Third, we evaluated to what extent AI-*segment* reproduces the analysis reported by radiologists. To this aim, we first compared the binary decision 'presence or absence of a lesion type' of *AI-segment* to the radiologist report considered as ground truth. A lesion type was detected by *AI-segment* when its estimated volumetry, averaged over both lungs, was above a certain threshold. The difference was then evaluated in terms of detection accuracy and F1 score, for two threshold values, using all scans of the IGR

dataset (Supp Table 1). Then, we compared disease extent as evaluated by radiologists to the one predicted by *AI-segment* (Supp Fig 5).

Machine learning models for severity classification based of CT scans (*AI-severity)*

The *AI-severity* model was defined as an ensemble of two sub-models, as illustrated in Supp Fig 4. Each sub-model predicted disease severity from CT scans without using any expert annotations at the slice level. Preprocessing of the data consisted of resizing the CT scans to a fixed pixel spacing of (0.7mm, 0.7mm, 10mm) and applying a specific windowing on the HU intensities. Each sub-model is composed of two blocks: a deep neural network called *feature extractor* and a penalized logistic regression. The two sub-models feature extractors are an EfficientNet-B0[59] pre-trained on the ImageNet public database and a ResNet50[60] pre-trained with MoCo v2[61] on one million CT scan slices from both Deep Lesion[62] and LIDC[63]. Each of these networks provide an embedding of the slices of the input CT scans into a lower-dimensional feature space (1280 for EfficientNet-B0 and 2048 for ResNet50). For the ResNet50-based sub-model, we reduced the dimension of the feature space using a principal component analysis with 40 components before applying logistic regression. A different windowing was applied on the CT scans before the feature extractor : (-1000 HU, 600 HU) for EfficientNet-B0 and (-1000 HU, 0 HU), (0 HU, 1000 HU) and (-1000 HU, 4000 HU) for ResNet50. Predictions of *AI-severity* were obtained by averaging predictions of the submodels using equal weights. Optimisation of the architecture of the network (preprocessing, feature extraction or model architecture and training, feature engineering, model aggregation) was performed using a 5 fold cross validation on the training set of 646 patients from KB.

CT scans may contain devices such as catheters (EKG monitoring, oxygenation tubing...) that are easily detectable in a CT and can bias prediction of severity (*i.e.* detecting the presence of a technical device associated with severity instead of detecting the radiological features associated with severity). In order to ensure that medical devices do not affect feature extraction, all voxels outside of the lungs were masked using a pre-trained U-Net lung segmentation algorithm [64].

Multivariate models to predict severity

The different models that combine multiple features to predict severity were fitted using logistic regression (*AI-segment*, trimodal *AI-segment*, *report*, trimodal *report/ScanCov*, trimodal *AI-severity*). Models were trained using cross validation with 5 folds on the training dataset of 646 patients from KB, and folds were stratified by age and severity outcome. Variables that were available for less than 300 patients of the training set (conjugated bilirubin and alanine) were not used. For the remaining variables, missing values were imputed by the average over patients of the training set. L2 regularization was applied when fitting logistic regression. The regularization coefficient was determined by maximizing the average AUC over the 5 cross-validation folds, using a range of different values ranging from 0.01 to 100. XGBoost algorithm was also evaluated but did not show superior performance on this dataset. We use pandas and scikit-learn to manipulate data, train and evaluate machine learning algorithms[65].

To select variables in the multivariate models, we considered a forward feature selection technique (Supp Fig 6). The first variable included in the model is the variable which provides the largest AUC values. Then, we computed AUC values for all models with two variables including the first one that has already been included. We continued this procedure until all variables were included. Performances of the models increased quickly when the first variables were included and then AUC values reached a plateau (Supp Fig 6). We used the elbow method to select the parsimonious set of variables that is found when a plateau of AUC is reached.

Other scores to predict severity and mortality

We performed a comparison with several multivariate scores of COVID severity or death. The COVID-GRAM score was the only multimodal score we considered that also includes information from CT scan[6]. When computing COVID-GRAM score, we assume that patients were not unconscious at admission and did not have hemoptysis as a symptom because these two information were missing from our dataset. The other scores we considered are based on clinical variables and possibly biological variables. They include the CALL (severity) score with clinical and biological information[9] as well as two other scores (the Yan et al. model for mortality prediction and the Colombi et al. model for severity prediction) that include clinical and biological information[7,15]. In order to compute the Yan et al. score, we considered the same features as the ones used by the authors and reproduced their training of an XGBoost model with a single tree and a maximum depth of 2[7]. We also considered the *MIT Covid Analytics* as a risk score for mortality (https://www.covidanalytics.io/mortality_calculator) and the CURB65 score developed to predict mortality for community-acquired pneumonia[66].

Data Availability

The dataset of patients hospitalized at Kremlin-Bicêtre (KB) and Institut Gustave Roussy (IGR) are stored on a server at Institut Gustave Roussy (IGR). The data are available from the first author upon request subject to ethical review.

Code Availability

Code to execute all the models presented in this article, including *ScanCov* score, *AI-segment* and *AI-severity* is available online on a public github repository.

Acknowledgements

We would like to thank J.-Y. Berthou, H. Berry, and Ph. Gesnouin from Inria and M. He, R. Patel, G. Rouzaud, B. Schmauch, J. Du Terrail from Owkin and F Lion from Gustave Roussy for their support.

Author Contributions

N.L., S.A., E.C.,P.H.,R.M.,N.L.,P.T., E.B.,M.S., A.S., F.C.,S.J., M.S., I.B., J.D.,JC.P., H.T.,E.P.,G.W., T.C., F.B.,MF.B.,M.B conceived the idea of this paper

N.L., S.A., E.C., H.G.,P.H., M.D., S.S., O.M., MP.T., JP.L.,R.M.,N.L.,P.T., E.B.,G.G, C.B.,S.J., F.G.,N.T.,Y.L., T.D., K.G., A.N., M.T., S.V., M.S., I.B., Y.B, E.P., M.A., J.D.,F.B., A.G.,J.D.,JC.P., H.T.,E.P.,G.W., T.C., F.B.,MF.B.,M.B participated to the acquisition and treatment of data

N.L., S.A., E.C.,P.H.,R.M.,N.L.,P.T., E.B.,S.J., M.S., P.J., I.B., J.D.,JC.P.,H.T.,E.P.,G.W., T.C., MF.B.,M.B.implemented the analysis

N.L., S.A., E.C.,P.H.,R.M.,N.L.,P.T., E.B.,S.J., M.S., I.B., J.D.,JC.P., H.T.,E.P.,G.W., T.C., MF.B.,M.B.contributed to the writing of the manuscript

Competing Interests statement

The authors declare the following competing interests:

- Employment: Michael Blum, Paul Herent, Rémy Dubois, Nicolas Loiseau, Paul Trichelair, Etienne Bendjebbar, Simon Jégou, Meriem Sefta, Paul Jehanno, Fabien Brulport, Olivier Dehaene, Jean-Baptiste Schiratti, Kathryn Schutte, Elodie Pronier, Jocelyn Dachary, Adrian Gonzalez, employed by Owkin
- Co-founders of Owkin Inc : Thomas Clozel, Gilles Wainrib.

References

1. Myers, L. C., Parodi, S. M., Escobar, G. J. & Liu, V. X. Characteristics of Hospitalized Adults With COVID-19 in an Integrated Health Care System in California. *JAMA* (2020) doi:10.1001/jama.2020.7202.
2. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
3. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* (2020) doi:10.1001/jama.2020.6775.
4. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern. Med.* (2020) doi:10.1001/jamainternmed.2020.0994.
5. Phua, J. *et al.* Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *Lancet Respir Med* **8**, 506–517 (2020).
6. Liang, W. *et al.* Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA Intern. Med.* (2020) doi:10.1001/jamainternmed.2020.2033.
7. Yan, L. *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* **2**, 283–288 (2020).
8. Levy, T. J. *et al.* Development and Validation of a Survival Calculator for Hospitalized Patients with COVID-19. *medRxiv* (2020) doi:10.1101/2020.04.22.20075416.
9. Ji, D. *et al.* Prediction for Progression Risk in Patients with COVID-19 Pneumonia: the CALL Score. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa414.
10. Mejia-Vilet, J. M. *et al.* A Risk Score to Predict Admission to Intensive Care Unit in Patients With COVID-19: The ABC-GOALS Score. *medRxiv* (2020).
11. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
12. Xie, J. *et al.* Association Between Hypoxemia and Mortality in Patients With COVID-19. *Mayo Clin. Proc.* **95**, 1138–1147 (2020).
13. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
14. Lippi, G. & Plebani, M. Laboratory abnormalities in patients with COVID-2019 infection. *Clin. Chem. Lab. Med.* (2020) doi:10.1515/cclm-2020-0198.
15. Colombi, D. *et al.* Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology* 201433 (2020).
16. Zhang, K. *et al.* Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell* (2020) doi:10.1016/j.cell.2020.04.045.
17. Zhao, W., Zhong, Z., Xie, X., Yu, Q. & Liu, J. Relation Between Chest CT Findings and Clinical Conditions of Coronavirus Disease (COVID-19) Pneumonia: A Multicenter Study. *AJR Am. J. Roentgenol.* **214**, 1072–1077 (2020).
18. Taieb, E. *et al.* Prognostic value of visual quantification of lesion severity at initial chest CT in confirmed Covid-19 infection: a retrospective analysis on 216 patients. *medRxiv* (2020).
19. Wu, J. *et al.* Chest CT Findings in Patients With Coronavirus Disease 2019 and Its Relationship With Clinical Features. *Invest. Radiol.* **55**, 257–261 (2020).
20. Clinical management of severe acute respiratory infection when COVID-19 is

suspected. https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-(ncov)-infection-is-suspected.
21. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
22. Hara, K., Kataoka, H. & Satoh, Y. Learning spatio-temporal features with 3D residual networks for action recognition. *Proc. IEEE* (2017).
23. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
24. Williamson, E. *et al.* OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients. *MedRxiv* (2020).
25. Liang, W. *et al.* Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *Lancet Oncol.* **21**, 335–337 (2020).
26. Miyashita, H. *et al.* Do patients with cancer have a poorer prognosis of COVID-19? An experience in New York City. *Ann. Oncol.* (2020) doi:10.1016/j.annonc.2020.04.006.
27. Dai, M. *et al.* Patients with cancer appear more vulnerable to SARS-COV-2: a multi-center study during the COVID-19 outbreak. *Cancer Discov.* (2020) doi:10.1158/2159-8290.CD-20-0422.
28. Wang, B., Li, R., Lu, Z. & Huang, Y. Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging* **12**, 6049–6057 (2020).
29. Gupta, R., Ghosh, A., Singh, A. K. & Misra, A. Clinical considerations for patients with diabetes in times of COVID-19 epidemic. *Diabetes Metab. Syndr.* **14**, 211 (2020).
30. Henry, B. M. & Lippi, G. Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection. *Int. Urol. Nephrol.* 1–2 (2020).
31. Lippi, G., Wong, J. & Henry, B. M. Hypertension and its severity or mortality in Coronavirus Disease 2019 (COVID-19): a pooled analysis. *Pol Arch Intern Med* **130**, 304–309 (2020).
32. Wang, X. *et al.* Comorbid Chronic Diseases and Acute Organ Injuries Are Strongly Correlated with Disease Severity and Mortality among COVID-19 Patients: A Systemic Review and Meta-Analysis. *Research* **2020**, 2402961 (2020).
33. Sprung, C. L. *et al.* Adult ICU Triage During the Coronavirus Disease 2019 Pandemic: Who Will Live and Who Will Die? Recommendations to Improve Survival. *Crit. Care Med.* (2020) doi:10.1097/CCM.0000000000004410.
34. Li, K. *et al.* The Clinical and Chest CT Features Associated With Severe and Critical COVID-19 Pneumonia. *Invest. Radiol.* **55**, 327–331 (2020).
35. Du, R.-H. *et al.* Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur. Respir. J.* **55**, (2020).
36. Li, X. *et al.* Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J. Allergy Clin. Immunol.* (2020) doi:10.1016/j.jaci.2020.04.006.
37. Ruan, Q., Yang, K., Wang, W., Jiang, L. & Song, J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med.* **46**, 846–848 (2020).
38. Yuan, M., Yin, W., Tao, Z., Tan, W. & Hu, Y. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. *PLoS One* **15**, e0230548 (2020).
39. Zhang, R. *et al.* CT features of SARS-CoV-2 pneumonia according to clinical presentation: a retrospective analysis of 120 consecutive patients from Wuhan city. *Eur. Radiol.* (2020) doi:10.1007/s00330-020-06854-1.
40. Li, K. *et al.* CT image visual quantitative evaluation and clinical classification of

coronavirus disease (COVID-19). *Eur. Radiol.* (2020) doi:10.1007/s00330-020-06817-6.
41. Li, Y., Yang, Z., Ai, T., Wu, S. & Xia, L. Association of 'initial CT' findings with mortality in older patients with coronavirus disease 2019 (COVID-19). *Eur. Radiol.* (2020) doi:10.1007/s00330-020-06969-5.
42. Liu, K.-C. *et al.* CT manifestations of coronavirus disease-2019: A retrospective analysis of 73 cases by disease severity. *Eur. J. Radiol.* **126**, 108941 (2020).
43. Ye, Z., Zhang, Y., Wang, Y., Huang, Z. & Song, B. Chest CT manifestations of new coronavirus disease 2019 (COVID-19): a pictorial review. *Eur. Radiol.* (2020) doi:10.1007/s00330-020-06801-0.
44. Xu, Z. *et al.* Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med* **8**, 420–422 (2020).
45. Tian, S. *et al.* Pulmonary Pathology of Early-Phase 2019 Novel Coronavirus (COVID-19) Pneumonia in Two Patients With Lung Cancer. *J. Thorac. Oncol.* **15**, 700–704 (2020).
46. Hagiwara, A., Fujita, S., Ohno, Y. & Aoki, S. Variability and Standardization of Quantitative Imaging: Monoparametric to Multiparametric Quantification, Radiomics, and Artificial Intelligence. *Invest. Radiol.* (2020) doi:10.1097/RLI.0000000000000666.
47. Wang, K. *et al.* Imaging manifestations and diagnostic value of chest CT of coronavirus disease 2019 (COVID-19) in the Xiaogan area. *Clin. Radiol.* **75**, 341–347 (2020).
48. Xiong, Y. *et al.* Clinical and high-resolution CT features of the COVID-19 infection: comparison of the initial and follow-up changes. *Invest. Radiol.* (2020).
49. Yang, R. *et al.* Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. *Radiology: Cardiothoracic Imaging* **2**, e200047 (2020).
50. Simpson, S. *et al.* Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. *Radiology: Cardiothoracic Imaging* **2**, e200152 (2020).
51. Hansell, D. M. *et al.* Fleischner Society: glossary of terms for thoracic imaging. *Radiology* **246**, 697–722 (2008).
52. La société d'Imagerie Thoracique propose un compte-rendu structuré de scanner thoracique pour les patients suspects de COVID-19. *SFR e-Bulletin* https://ebulletin.radiologie.fr/actualites-covid-19/societe-dimagerie-thoracique-propose-compte-rendu-structure-scanner-thoracique (2020).
53. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
54. Hara, K., Kataoka, H. & Satoh, Y. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* 3154–3160 (2017).
55. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* 234–241 (Springer International Publishing, 2015).
56. Chen, S., Ma, K. & Zheng, Y. Med3D: Transfer Learning for 3D Medical Image Analysis. *arXiv [cs.CV]* (2019).
57. Yang, J. *et al.* Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Med. Phys.* **45**, 4568–4581 (2018).
58. Baldeon Calisto, M. & Lai-Yuen, S. K. AdaEn-Net: An ensemble of adaptive 2D-3D Fully Convolutional Networks for medical image segmentation. *Neural Netw.* **126**, 76–94 (2020).
59. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv [cs.LG]* (2019).
60. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition.

arXiv [cs.CV] (2015).
61. Chen, X., Fan, H., Girshick, R. & He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv [cs.CV]* (2020).
62. Yan, K., Wang, X., Lu, L. & Summers, R. M. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imaging (Bellingham)* **5**, 036501 (2018).
63. LIDC-IDRI - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.
64. Hofmanninger, J. *et al.* Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem. *arXiv [eess.IV]* (2020).
65. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
66. Lim, W. S. *et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* **58**, 377–382 (2003).