AI-based multi-modal integration 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[2], 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

With 15% of severe cases among hospitalized patients[1], the SARS-COV-2 pandemic has put tremendous pressure on Intensive Care Units, and made the identification of early predictors of future severity a public health priority. We collected clinical and biological data, as well as CT scan images and radiology reports from 1,003 coronavirus-infected patients from two French hospitals. Radiologists' manual CT annotations were also available. We first identified 11 clinical variables and 3 types of radiologist-reported features significantly associated with prognosis. Next, focusing on the CT images, we trained deep learning models to automatically segment the scans and reproduce radiologists' annotations. We also built CT image-based deep learning models that predicted future severity better than models based on the radiologists' scan reports. Finally, we showed that including CT scan features alongside the clinical and biological data yielded more accurate predictions than using clinical and biological data alone. These findings show that CT scans provide insightful early predictors of future severity.

Previous studies have demonstrated that risk factors for severe evolution include demographic variables such as age, comorbidities, and biological variables measured within 2 days of patient admission[2–4]. Beyond clinical and biological variables, computerized tomography (CT) scans are also potential sources of information: the degree of pulmonary inflammation is associated with clinical symptoms and severity[5,6], and the extent of lung abnormality is predictive of severe disease evolution[7,8]. Here we evaluated to what extent visual or AI-based analysis of CT scans at patient admission added information about future severe disease evolution once clinical and biological data had been taken into account.

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. Additionally, 292 CT scans were later annotated manually by radiologists (see supplementary materials). Summary statistics for the clinical, biological, and CT scan data are provided in Figure 1.

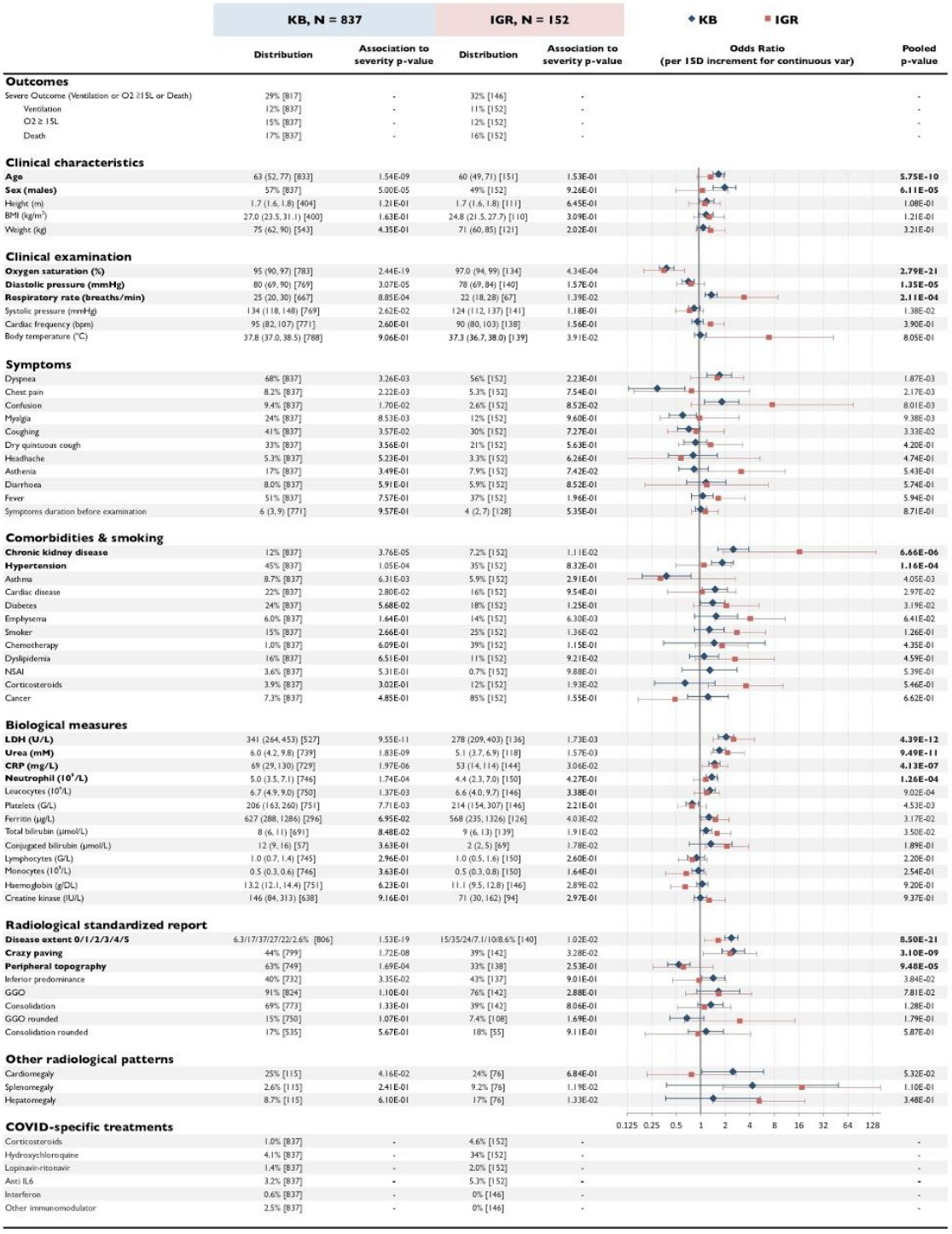


Figure 1: Population description for the KB and IGR hospitals. Among the 1,003 patients of the study, biological and clinical variables were available for 989 individuals. Categorical variables are expressed as percentages [available]. Continuous variables are shown as median (IQR) [available]. Association with severity are reported with p-values for each center and the pooled p-value has been obtained with Stouffer's method to combine p-values. p-values that are shown are not adjusted for multiplicity. Variables and pooled p-values are in bold when the variable is significant after Bonferroni adjustment to account for multiple testing across the 63 variables. For continuous variables, odds ratios are computed for an increase of one standard deviation of the continuous variable. KB odds ratios are in blue, IGR in red.

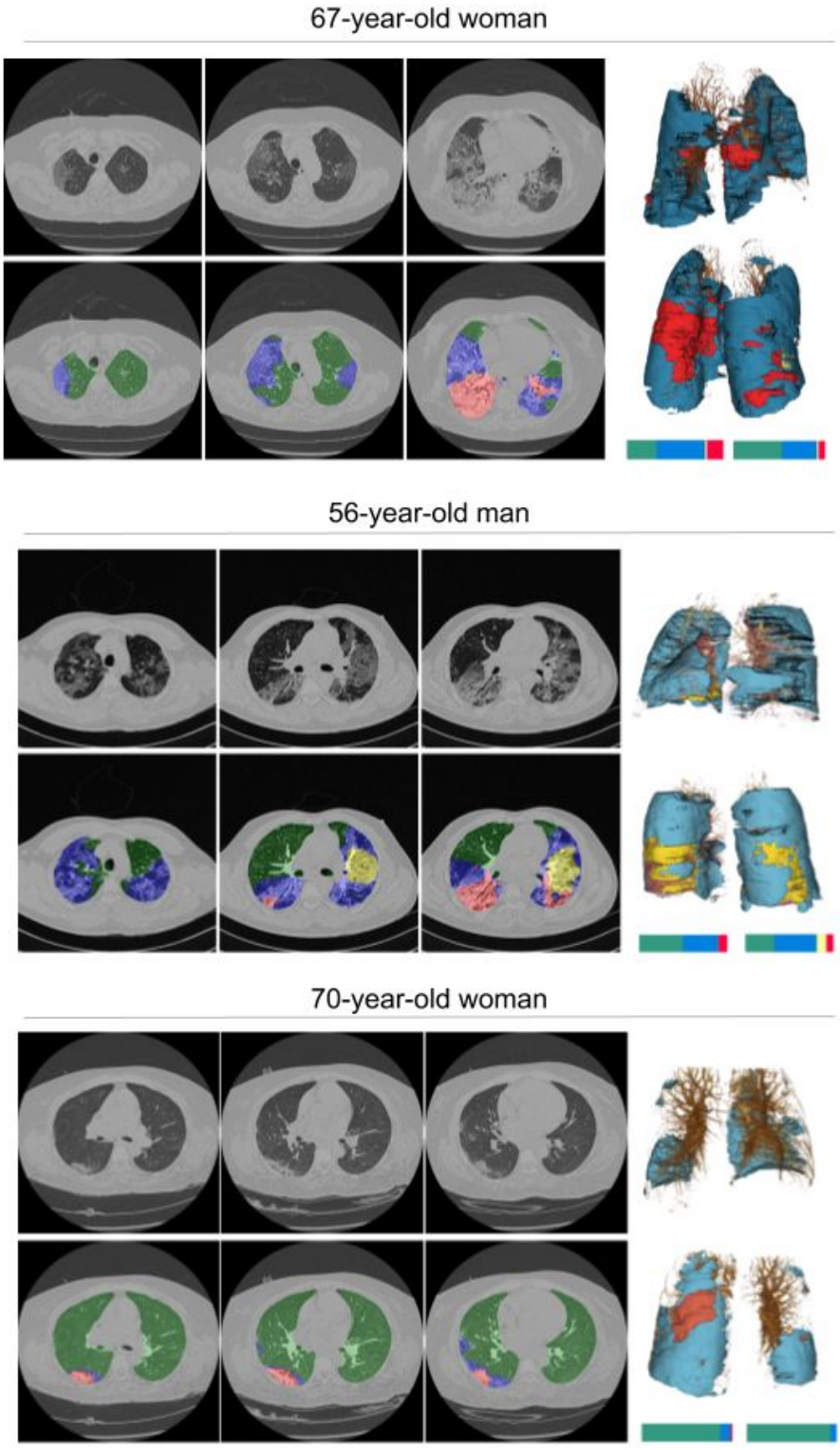


Figure 2: **Axial chest CT scans and segmentation results** COVID-19 radiology patterns, as provided by AI-segment, for 3 patients with COVID-19. Green/transparent: sane lung; blue: GGO; yellow : crazy paving; red: consolidation. (Top) 67-year-old woman with diffuse distribution, and multiple large regions 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).

Coronavirus progression is evaluated by the World Health Organization on a 1 to 10 scale, severe scores of 5 or more corresponding to an oxygen flow rate of 15 L/min or higher, or the need for mechanical ventilation, or patient death[9]. We first evaluated how clinical and biological variables measured at admission were associated with future severe progression (score of 5 or more). These variables were available for 989 individuals, and we computed the severity odds ratios for each individual variable, and at each hospital center (Figure 1). When combining association results from the two centers, we found 11 variables significantly associated with severity (P <0.05/63 to account for testing 63 variables, Figure 1): **age** (Odds Ratio [OR] KB 1.66 (1.41-1.96), OR IGR 1.04 (0.50-2.15), OR IGR 1.32 (0.90-1.93), $P_{Stouffer}$ = 5.75e-10), **sex** (OR KB 1.95 (1.41-2.69), OR IGR 1.04 (0.50-2.15), $P_{Stouffer}$ = 6.10e-05), **hypertension** (OR KB 1.84 (1.35-2.51), OR IGR 1.09 (0.50-2.36), $P_{Stouffer}$ = 1.15e-04), **chronic kidney disease** (OR KB 2.51 (1.62-3.69), OR IGR 16.29 (1.89-140.12), $P_{Stouffer}$ = 6.66e-06), **respiratory rate** (OR KB 1.34 (1.13-1.59), OR IGR 3.37 (1.28-8.86), $P_{Stouffer}$ = 2.10e-04), **oxygen saturation** (OR KB 0.38 (0.31-0.47), OR IGR 0.35 (0.20-0.63), $P_{Stouffer}$ = 2.79e-21), **diastolic pressure** (OR KB 0.70 (0.53-0.83), OR IGR 0.76 (0.51-1.11), $P_{Stouffer}$ = 1.35e-05), **CRP** (OR KB 1.47 (1.25-1.72), OR IGR 1.50 (1.04-2.16), $P_{Stouffer}$ = 4.13e-07), **LDH** (OR KB 2.05 (1.65-2.54), OR IGR 2.53 (1.42-4.53), $P_{Stouffer}$ = 4.38e-12), **polynuclear neutrophil** (OR KB 1.36 (1.13-1.60), OR IGR 1.15 (0.81-1.64), $P_{Stouffer}$ = 1.25e-04), and **urea** (OR KB 1.70 (1.43-2.01), OR IGR 2.13 (1.33-2.42), $P_{Stouffer}$ = 9.49e-11). This confirms the literature reported prognostic value of these 11 clinical and biological markers.[2,4,10–14]

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 1.64 (1.12-2.38), $P_{Stouffer}$ = 8.50e-21) and (ii) **crazy paving** *(*OR KB 2.50 (1.82-3.44), OR IGR 2.28 (1.07-4.88), $P_{Stouffer}$ = 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 0.61 (0.26-1.42), $P_{Stouffer}$ = 9.47e-05). This confirms the reported negative impact of disease extent[7,15,16]. We hypothesize that peripheral topography has a positive impact on prognosis because peripheral lesions could be less extended.

We next trained a deep neural network called *AI-segment* (Supp Figure 1) to segment radiological patterns and provide automatic quantification [18,19] 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[17]: ground glass opacity or GGO, crazy paving, and finally consolidation. *AI-segment* was trained on 161 patients from KB and evaluated on 132 patients from IGR, of which 14 fully annotated, and 118 partially annotated. The mean absolute error in volume prediction for the fully annotated scans was 6.94% for GGO, 1.01% for consolidation, and 7.21% for sane lung (no crazy paving was present in these scans). On the larger cohort of partially annotated scans, the accuracy with respect to the radiologist score was 78% for GGO, 67% for crazy paving, and 74% for consolidation (for a 1% detection threshold on the *AI-segment* result, Supp Table 1). *AI-segment* also accurately quantified the disease extent (Supp Figure 3). *AI-segment* visual results were also consistent with radiologist observations (See Figure 2 for three representative cases). We lastly evaluated to what extent the *AI-segment* trained on CT

scans provided finer information about future severity compared to radiologists' scan reports. Using predicted volumes from *AI-segment*, we found that GGO (OR KB 1.8 (1.5-2.16), OR 1.7(1.18-2.43), $P_{Stouffer}$ = 3.45e-11), crazy paving (OR KB 1.57 (1.26-1.97), OR IGR 1.38 (0.95-1.99), $P_{Stouffer}$ = 7.27e-05) consolidation (OR KB 1.86 (1.53-2.25), OR IGR 1.87 (1.26-2.77), $P_{Stouffer}$ = 1.43e-11) and extent of disease (OR KB 2.14 (1.77-2.6), OR IGR 1.87 (1.28-2.73), $P_{Stouffer}$ = 3.13e-16) were all associated with severity (accounting for multiple testing). This confirms that automatic estimation of lesion volumes can add more precise measures of future severity to the radiologists' scan reports (Supp Table 2) [8] .

We next evaluated the prognostic value of CT scans alone through three different models. The first model called *report* included variables from the radiological report only. The second was based on the automatic lesion volumes measured by *AI-segment*. The third called *AI-severity* used a weakly supervised approach with no radiologist-provided annotations (Supp Figure 2)[20]. All three models were trained on 646 KB patients, tested on 150 KB validation patients, and validated on the independent IGR dataset of 137 patients (Figure 3). On the validation set from KB hospital, *report* was outperformed by *AI-severity* but not by *AI-segment* ($AUC_{AI\text{-}severity}$ = 0.76, $AUC_{AI\text{-}segment}$ = 0.68, AUC_{report} = 0.72). On the independent IGR validation set, both *AI-segment* and *AI-severity* outperformed the model *report* ($AUC_{AI\text{-}severity}$ = 0.70, $AUC_{AI\text{-}segment}$ =0.68, AUC_{report} =0.66). Our follow up analyses revealed that the predictive performance of *AI-severity* was strong in part because the internal representation of the neural network captures clinical features from the lung CTs, such as age, on top of the known COVID-19 radiology features (see interpretability of *AI-severity* in Supp Material).

Lastly, we evaluated whether CT scans have prognostic value beyond what can be inferred from clinical and biological characteristics alone. We therefore compared the performance of trimodal CT scan / clinical / biological models to bimodal clinical / biological models. We compared model performances for three outcomes: our 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 studied in the literature, "death or ICU admission", and "death". We built a trimodal version of *report*, *AI-segment*, and *AI-severity*, adding clinical and biological information to the original CT scan-based models by implementing a greedy search approach to include optimal variables (Supp Figure 4). All three trimodal models performed consistently better than the bimodal biological/clinical model (Figure 3 and Supp Table 3), whether it be trimodal *report*, *AI-segment*, or *AI-severity (*mean AUC increase of 0.02-0.03*)*. They also outperformed clinical/biological models from literature (Colombi at al model [7] and MIT COVID analytics model). Of note, the fact that the models trained with patients from the KB hospital had good performances when evaluated on IGR hospital is evidence of their robustness, especially since these two hospitals receive patients with very different comorbidities (85% of cancer patients at IGR and 7% at KB). Taken together, these consistent results confirm the added prognostic value of CT scans. Importantly, while trimodal *AI-severity* generally outperformed trimodal *report* across all outcomes, and trimodal *AI-segment* sometimes outperformed *report*, the AUC difference was always modest (max increase of 0.03 for *AI-severity* vs *report*, and max increase of 0.02 for *AI-segment* vs *report*), showing that the incorporation of CT-scan analyses, no matter what the method, is the strongest performance booster. Therefore beyond AI modeling, our study shows that a

composite scoring system integrating selected radiological measurements with key clinical and biological variables provides accurate predictions and can rapidly become a reference scoring approach for severity prediction.

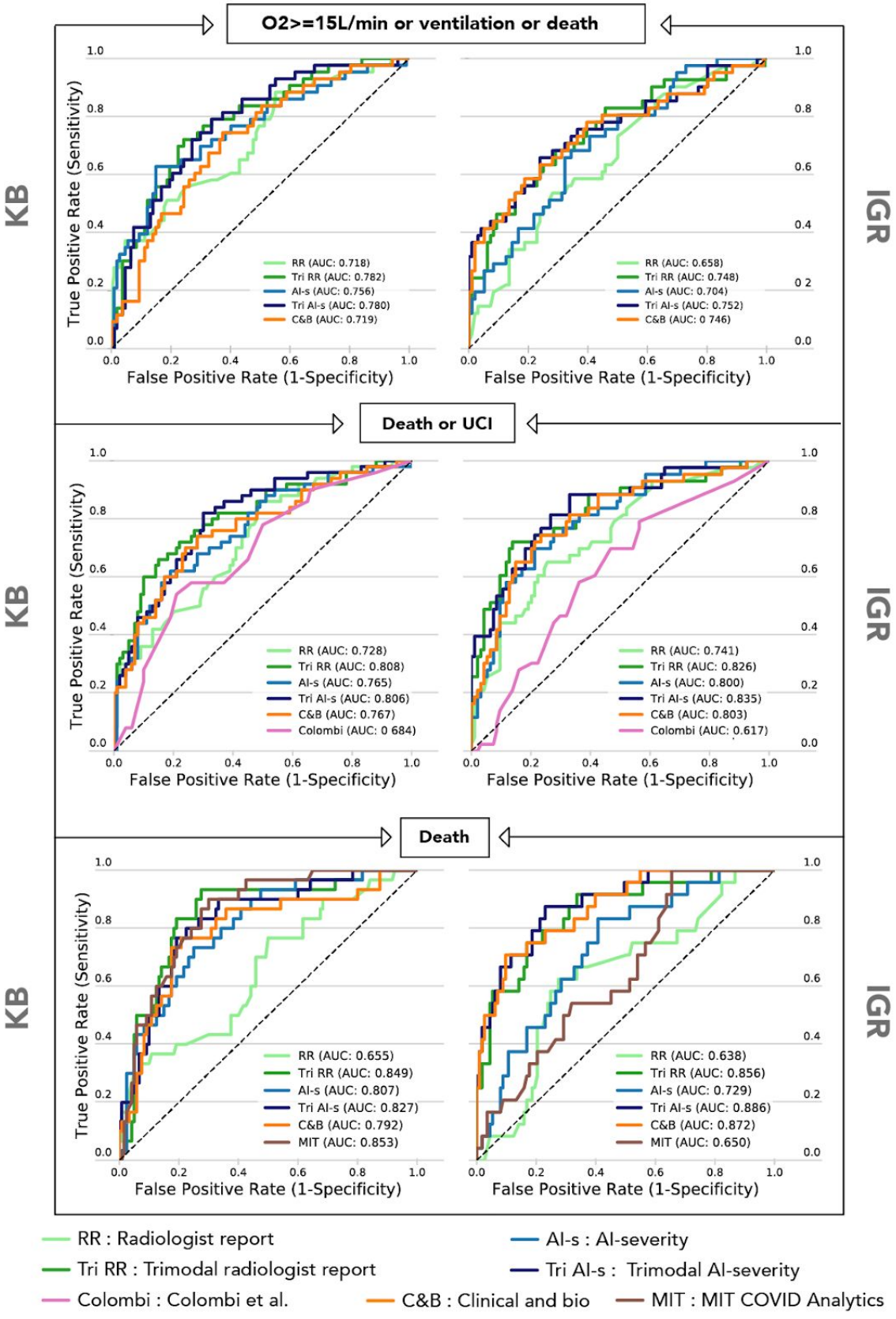


Figure 3: Receiver operating characteristic (ROC) curves of the models that predict severity. Models were evaluated on two distinct validation sets consisting of 150 patients from KB (left panels) and 137 patients from IGR (right panels).

Our retrospective study conducted on two French hospitals shows that future disease severity markers are present within routine CT scans performed at admission, and these can be identified and quantified via AI-based scoring, providing useful and interpretable elements for prognosis.

Acknowledgements

We would like to thank J.-Y. Berthou, H. Berry, and Ph. Gesnouin from Inria and B. Schmauch, G. Rouzaud, and R. Patel from Owkin 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. Guan, W.-J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
2. 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).
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)
4. Yang, J. *et al.* Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. *Int. J. Infect. Dis.* **94**, 91–95 (2020).
5. 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).
6. 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).
7. Colombi, D. *et al.* Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology* 201433 (2020).
8. Zhang, K. *et al.* Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell* (2020).
9. 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.
10. Livingston, E. & Bucher, K. Coronavirus Disease 2019 (COVID-19) in Italy. *JAMA* (2020)
11. 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)

12. 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).
13. Huang, R. *et al.* Clinical Findings of Patients with Coronavirus Disease 2019 in Jiangsu Province, China: A Retrospective, Multi-Center Study. *PLoS Negl. Trop. Dis.* 14: e0008280. (2020)
14. Wu, Z. & McGoogan, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* (2020)
15. 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).
16. 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)
17. Wang, Y. *et al.* Temporal Changes of CT Findings in 90 Patients with COVID-19 Pneumonia: A Longitudinal Study. *Radiology* 200843 (2020).
18. 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).
19. Hara, K., Kataoka, H. & Satoh, Y. Learning spatio-temporal features with 3D residual networks for action recognition. *Proc. IEEE* (2017).
20. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
21. 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)

Supplementary material of "AI-based *multi-modal* integration of clinical characteristics, lab tests and chest CTs improves COVID-19 outcome prediction of *hospitalized* patients"

Description of the retrospective study

Data were collected at two French hospitals (Kremlin Bicêtre Hospital (KB), APHP, Paris, and Gustave Roussy Hospital (GR), Villejuif). CT scans, clinical, and biological data were collected in the first 2 days after hospital admission.

This study has received the approval of both hospitals ethic committees and we submit 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. Also an information letter was sent to all patients included in 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[1]. 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 from the clinical examination include: sex, age, 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.

Chest Thoracic (CT) imaging

CT scan acquisition

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 the 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), 1mm 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.

The scans performed were independently examined by experienced radiologists using a standard workstation in the clinical image archiving and transmission system. All radiologists were informed of patients clinical status (suspicion of COVID-19, clinical signs of severity).

Definition of CT Features

COVID-19 associated CT imaging features identified by radiologists were defined following ACR recommendation[1]. The term parenchymal opacification is applied to any homogeneous increase in lung density on chest CT. When this parenchymal opacification is dense enough to obscure the vessels margins and airway walls and other parenchymal structures, it is called consolidation. Ground-glass attenuation is defined as an increase in lung density not sufficient to obscure vessels or preservation of bronchial and vascular margins crazy-paving pattern was defined as ground-glass opacification with associated interlobular septal thickening[2] .

For 959 patients, CT imaging characteristics were evaluated and the following findings were reported: ground glass opacity (rounded / non rounded / absent), consolidation (rounded / non rounded / absent) interlobular septal thickening or “crazy paving” (present / absent), subpleural line, lymph node enlargement, pleural effusion, and pericardial effusion, according to morphological descriptors based on recommendations of the Fleischner Nomenclature Committee[2].

The results of the CT were examined in terms of location, distribution, size and type. The location refers to the different lobes and segments involved (lower or medium or upper). The distribution was described as peripheral (1/3 external of the lung), central (2/3 internal), or both central and peripheral.

The assessment of the size and extent of lung involvement was based on a visual classification of lung anatomy according to the evaluation criteria established by the French Society of Radiology (SFR)[3]. The size of the lesion was assessed; the volume of the lung affected 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.

Automatic extraction from radiological report

Radiological features from radiological reports were automatically extracted using Optical Character Recognition and regular expression functions.

Annotation scenario of CT scans by radiologists in order to train the AI-Volumetry model

Two radiologists (4 and 9 years of experience) examined and annotated 292 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. One AI and imaging PhD student provided full 3D annotation of the four classes on 22 anonymized chest scans using the 3D Slicer software.

The presence of organomegaly was also notified when present, as a binary class. When multiple CT images were available for a single patient, the scan to analyze was selected using the SPYD software.

Machine learning models

Models for segmentation of CT scans (*AI-segment*)

In the proposed pipeline called AI-*segment* for lesion segmentation from CT scans, we deployed 3 segmentation networks: 3D Resnet50[4] , 2.5D U-Net, and 2D U-Net [5]. These are three powerful convolutional neural networks that have achieved state of the art performance in numerous medical image segmentation tasks. 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 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 292 scans (153 from KB and 118 from IGR)

2D U-Net was trained for left/right lung segmentation while 3D ResNet and 2.5D U-Net were used for lesion segmentation. 3D ResNet50 was trained on 8 KB FAS. We used Stochastic Gradient Descent for parameter optimization and a learning rate starting of 0.1 with a decay factor of 0.1 every 20 epochs. The network was trained for a total of 100 epochs. As for 2.5D U-Net, Adam optimization algorithm was used with learning rate, weight decay, gradient clipping and learning rate decay parameters set respectively to 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 3D resnet50, 153 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 (on average 16 slices are annotated per PAS) in the scans which would not satisfy the volumetric requirements of the 3D ResNet50 input. Finally, for the left/right lung segmentation, the 2DU-Net was trained on the 8 KB FAS. Similarly to 2.5D U-Net, Adam optimization algorithm was used with learning rate, weight decay, gradient clipping and learning rate decay parameters set respectively to 1e-3, 1e-8, 1e-1, and 0.1 at epoch 70 over 104 epochs. Both 2.5D U-Net and 2D U-Net use affine transformation and contrast change for data augmentation while 3D resnet50 uses affine transformation, contrast change, thin plate splines, and flipping. 3D ResNet and 2.5D U-Net are trained through the minimization of the cross entropy loss and 2D U-Net minimizes the binary cross entropy loss. All training was performed on NVIDIA Tesla V100 GPUs and Pytorch is the used framework. During the validation phase, ensemble inference[6] is performed on all the available scans.

Models for severity classification based of CT scans (*AI-severity)*

The *AI-severity* model is defined as an ensemble of four sub-models, as illustrated in Supp Fig 2. Each of these sub-models is designed to predict the disease severity from CT scans. Since they do not require expert annotations at the slice level, these sub-models fall in the scope of *weakly supervised learning*. The preprocessing of the data consisted in resizing the CT scans to 10mm pixel spacing along the vertical axis and obtaining a segmentation of the lungs using a pre-trained U-Net algorithm[7]. Each sub-model is composed of two blocks: a deep neural network called *feature extractor* and a logistic regression. CT scans may contain biases such as catheters (EKG monitoring, oxygenation tubing...) that are easily detectable in a CT and can bias the prediction of severity (*i.e.* predict the presence of a technical device associated with severity instead of predicting the radiological features associated with severity). In order to ensure that these biases do not affect the features, the lung segmentation mask was applied before the features were extracted. As a result, only the lungs were visible to the *feature extractor*.

Two of the sub-models used an EfficientNet-B0[8] pre-trained on the ImageNet public database as feature extractor while the other two used a ResNet50[9] pre-trained with MoCo v2[10] on one million CT scan slices from both Deep Lesion[11] and LIDC[12]. Each of these networks provide an embedding of the slices of the input CT scans into a lower-dimensional (1280 for EfficientNet-B0 and 2048 for ResNet50 with MoCo v2) feature space. A windowing used for selecting specific ranges of intensities was also applied on the CT scans before the features extraction. For the two sub-models based on the EfficientNet-B0, the image intensities were respectively clipped in the (-1000 HU, 200 HU) and (-1000 HU, 600 HU) range. For one of the remaining two sub-models (based on ResNet50 with MoCo v2), the (-1350 HU, 150 HU) range was used whereas for the last one, a combination of the following ranges was used: (-1000 HU, 0 HU), (0 HU, 1000 HU) and (-1000 HU, 4000 HU). Finally, for each of these sub-model, a Logistic Regression (with ridge penalty) was used to predict the disease severity from the averaged features. For the ResNet50-based sub-models, a Principal Component Analysis (PCA) with 40 components was used to reduce the dimensionality of the feature space before the Logistic Regression was applied. All the

sub-models were equally weighted in the ensemble and the disease severity predictions of the *AI-severity* model were obtained by averaging the prediction of the models in the ensemble.

Interpretability of *AI-severity*

An interpretability study was conducted on *AI-severity* to get a better understanding of its performances. The correlation between the internal representation of the sub-models (*i.e.* the input of the logistic regression),radiological and clinical variables were analyzed. By replacing the output of the logistic regression by variables from the radiology reports, AUC on the KB validation set of 150 patients were 94.1% for disease extent (threshold >2), 71.4% for crazy paving, 67.1% for condensation and 74.8% for GGO, showing that the feature extractors correctly captured part of the radiology signal. More interestingly, it was also possible to correlate internal representations with clinical variables such as age (AUC 85.1% with a threshold of 60 years old), sex (AUC 85.2%) or oxygen saturation (AUC 76.2%, threshold 90%). As a comparison, a logistic regression trained on the radiology report variables only gets respectively AUC scores of 70.0%, 59.9% and 67.8%. This gap shows that the *AI-severity* internal representations present within the neural network capture clinical information directly from CT scans.

Models for multimodal integration

The models used to predict the outcome from multiple modalities are logistic regressions, trained by cross validation with 5 folds on the training dataset of 646 patients from KB, stratified by age and outcome. Variables that were filled for less than 300 patients (conjugated bilirubin and alanine) were not used. For the remaining variables, missing values were simply replaced by the average over patients of the training set. L2 regularization was applied to the weights of the models. The regularization coefficient value was chosen by comparing the results obtained in cross validation with different values, ranging from 0.01 to 100. The value maximizing the average AUC over the 5 folds was selected. We use pandas and scikit-learn to manipulate data and perform machine learning algorithms [13].

Selection of clinical and biological variables added to the models based on CT scan variables

Clinical and biological variables were selected through a forward feature selection technique (Supp Fig 4). At baseline (left of the figure), a model was trained in cross-validation using only a fixed set of variables. Three initial sets were considered here: radiologist report, AI Lungs and AI volumetry. The variables encoded in the radiologist report includes a presence/absence coding of Ground Glass opacity (GGO), rounded GGO, Crazy paving, Consolidation, Consolidation rounded, Topography peripheral, and Predominance inferior, as well as disease extent, which is a semi automatic assessment of the amount of lesions in

the lung. The AI-Lung model includes the one variable output of the neural network model to predict severity and the AI volumetry model includes the automatic quantification of the ground glass, consolidation and crazy paving pattern, and the automatic quantification of disease extent. For comparison, the procedure was also performed starting from an empty set of variables (clinical only).
The added prognosis value of every clinical or biological variable was then assessed separately, by training a new model using this variable in addition to the previous set. The variable resulting in the largest AUC score was added to the selection. This procedure was repeated for 20 iterations. For every initial selection, performances of the models increased quickly at first (left part of Supp Fig 4), then reached a plateau (right half of the figure), indicating that the variables added after the tenth iteration did not significantly increase the predictive power of the models. Thus, for every case, only the ten best clinical and biological variables were selected.

Training and evaluation of models

To predict severity, models were trained on 646 patients from KB, which included the training set of *AI-segment*, and evaluated on two distinct evaluation sets, with 150 patients from KB and 137 patients from IGR. The prediction is performed using the logistic regression approach.

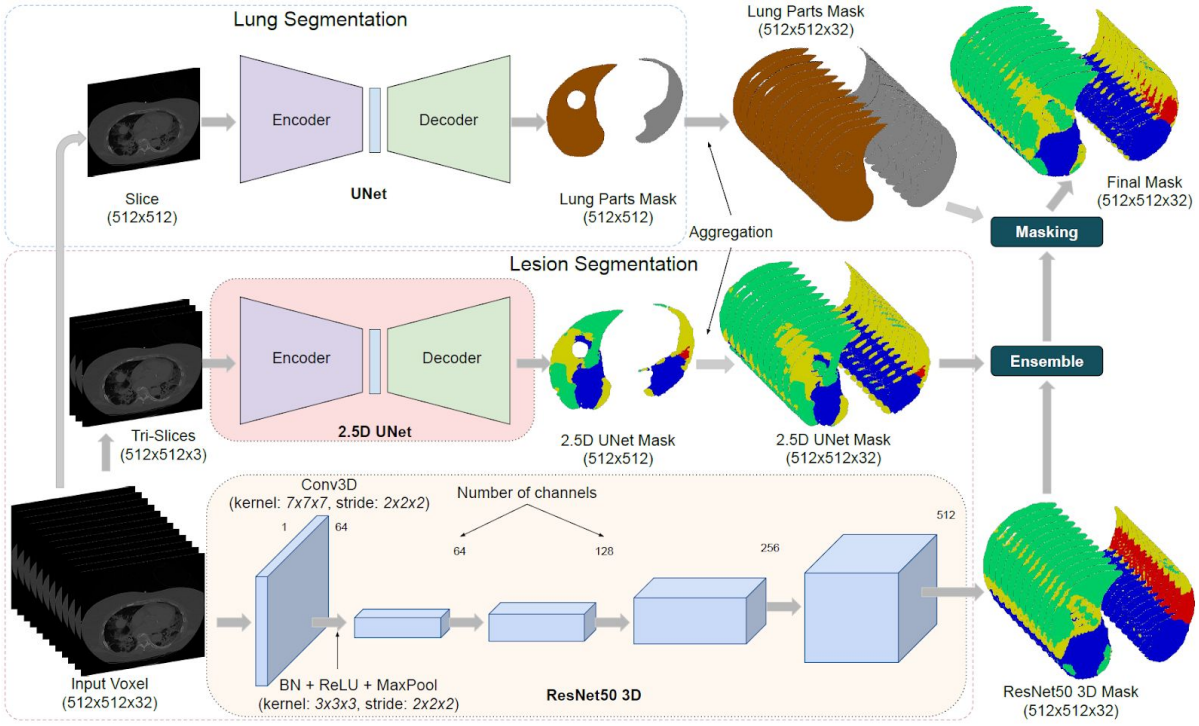
We evaluated models that predict severity using the Area Under the Curve (AUC) and differences between AUC values were tested using DeLong test [14].

We evaluated the segmentation model *AI-segment* using mean absolute error that is defined as the average, over the available fully annotated CT scans in the validation set, of the absolute value of the difference between the ground truth percentage of each lesion type (deduced from annotations) and the estimated ones. We also evaluated the detection accuracy per lesion with respect to the reported radiologist scores, defined as the percentage of correctly predicted classes by *AI-segment* (GGO ; CP ; Consolidation) among the validation set. A given lesion type, in the *AI-segment result,* is considered as present when the estimated volumetry of the lesion type, averaged over both lungs, is above a certain threshold (here, we reported results for threshold 1% and 2%).

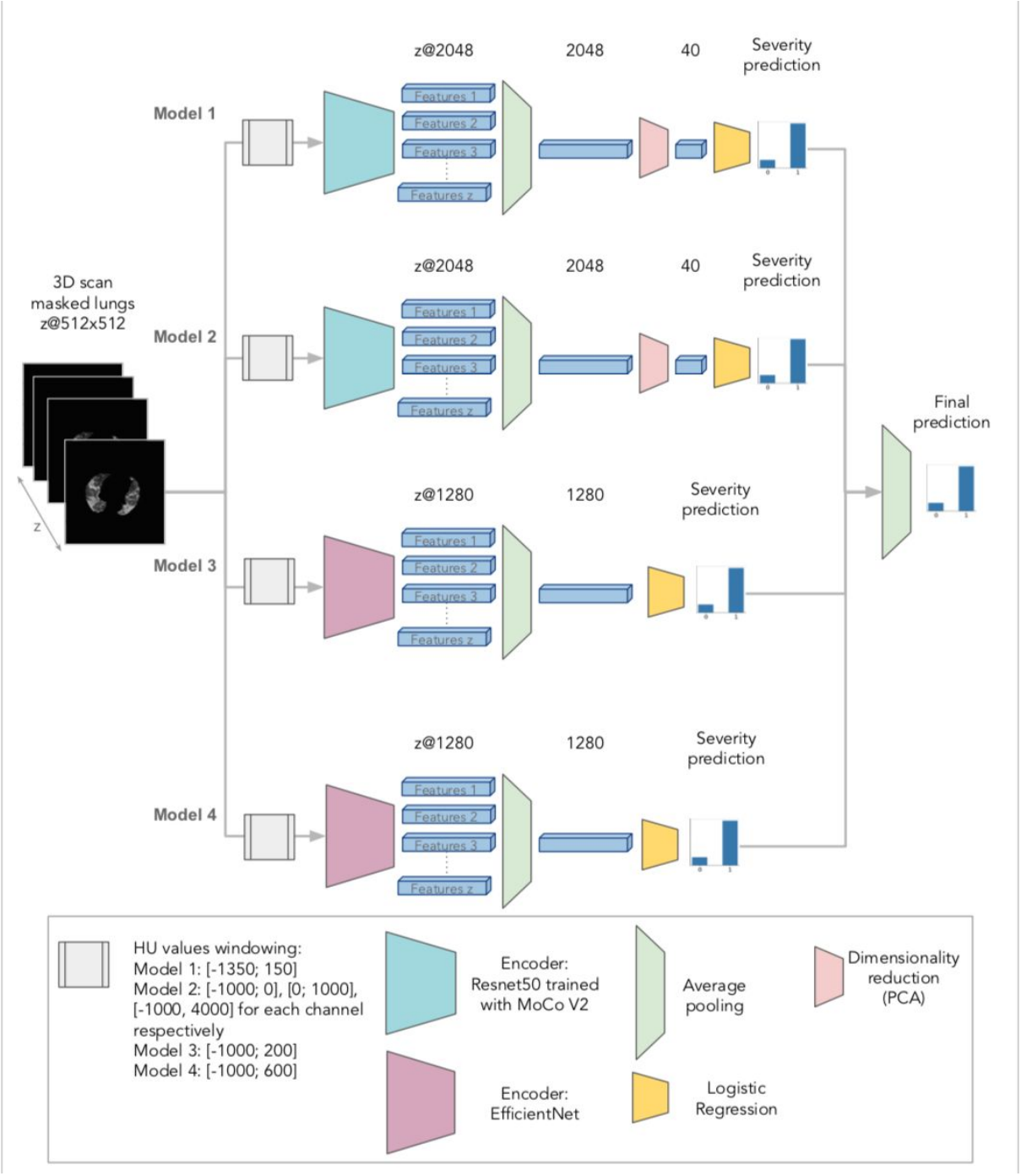
Benchmark models

We use the clinical and biological variables previously proposed in a multivariate risk score for severity, which is defined as admission to ICU or death, and we retrain a logistic regression model using these variables [15]. We also considered the *MIT Covid Analytics* calculator as a risk score for mortality (https://www.covidanalytics.io/mortality_calculator).

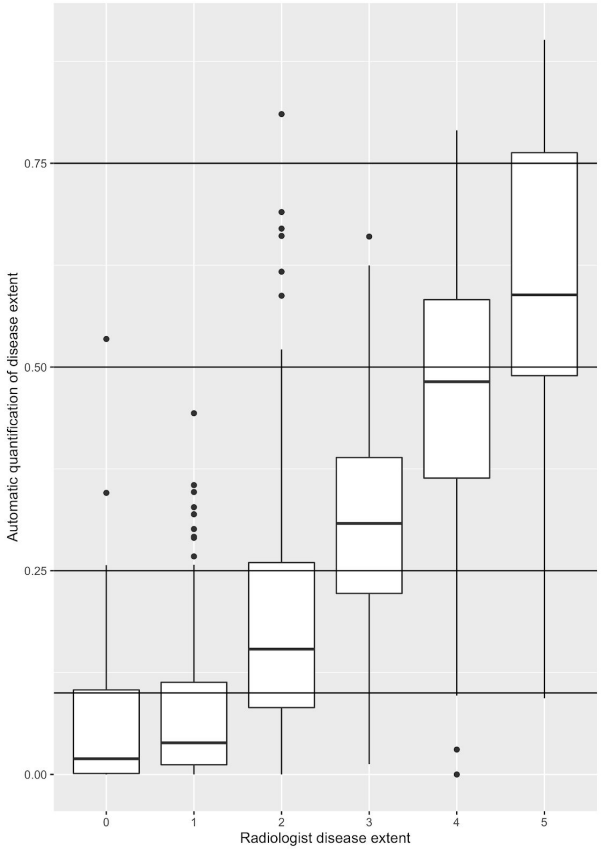
Supplementary Figures and Tables



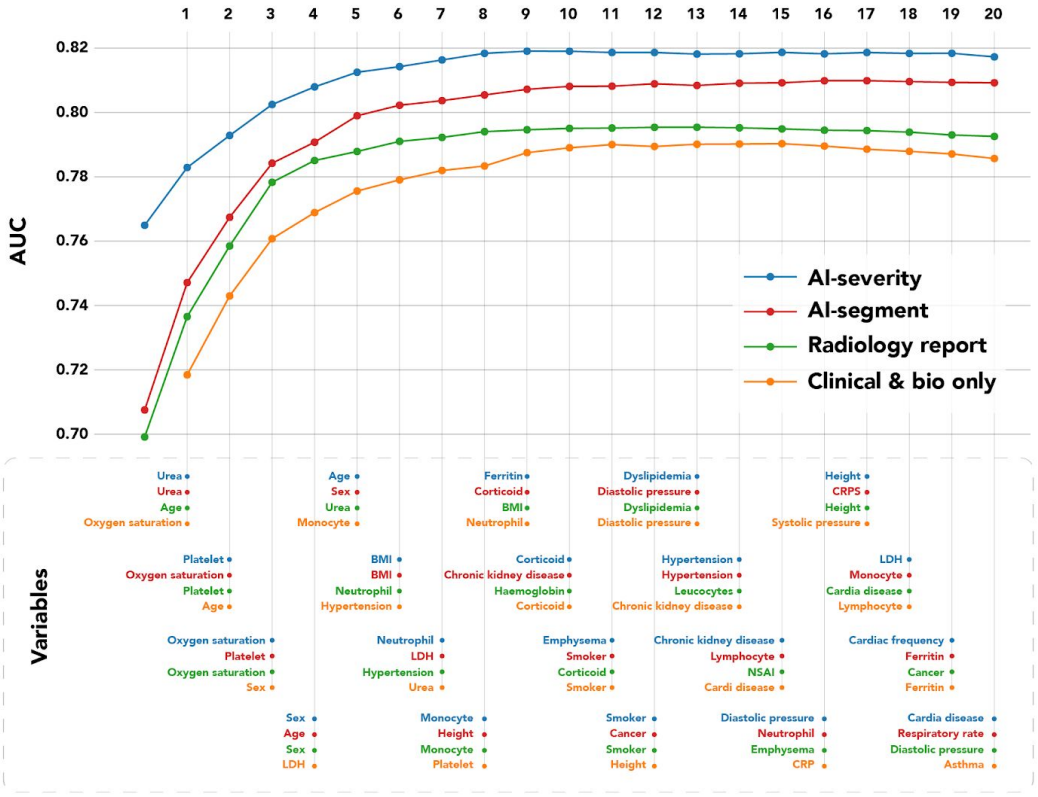
Supp Fig. 1: *AI-segment* architecture - Proposed pipeline to generate lesion volumetry estimates from patient CT scans employing ensemble of segmentation networks. Normalized patient scans are provided to our trained 2.5D U-Net and 3D ResNet50. The masks predicted from both models are then merged by geometric mean. In parallel, we segment left-right lungs from the patient scans using a dedicated U-Net. Finally, the left-right lung mask is used to mask-out lesions in left and right lungs from the ensemble output. This pipeline utilizes the complementary features learned by a weak model (2.5D U-Net) and a strong one (3D ResNet50).



Supp Fig. 2: *AI-severity* model to predict severity from 3D chest CT scans.
Two different pipelines were used: one using Resnet50 (trained with MocoV2 on 1 million public CT scan slices) as encoder (models 1 & 2) and one using EfficientNet B0 as encoder (models 3 & 4).



Supp Fig. 3: Boxplot of the automatic quantification of disease extent by *AI-segment* versus disease extent as estimated by a radiologist. The coding of the report is as follows: 0 (0% of lesions), 1 (<10% of lesions), 2 (between 10 and 25% of lesions), 3 (between 25 and 50% of lesions), 4 (between 50 and 75% of lesions), 5 (more than 75% of lesions).



Supp Fig. 4: AUC curve as a function of the number of clinical and biological information added to the multimodal model. Variables included in the models consist of CT scan variables only and then a greedy algorithm adds clinical or biological variables iteratively. At each step of the algorithm, the variable that results in the largest increase of AUC score is added.

	GGO	Crazy paving	Consolidation
Accuracy (1% threshold)	**0.7829**	0.6712	0.7419
Accuracy (2% threshold)	0.7679	0.6712	**0.7558**

Supp Table 1: detection accuracy computed for the binary decision “presence or not of a lesion type” for *AI-segment* (threshold of the predicted disease extent, maximum of both lungs), when compared to standardized radiologist report, on the IGR cohort.

Variable	Center	Odds ratio (95%lower - 95% upper)	P-value	P-value Stouffer
GGO AI	KB	1.8 (1.5-2.16)	2.86e-10	3.45e-11
GGO AI	IGR	1.7(1.18-2.43)	0.00424	
Crazy Paving AI	KB	1.57 (1.26-1.97)	6.37e-05	7.27e-05
Crazy Paving AI	IGR	1.38 (0.95-1.99)	0.08712	
Consolidation AI	KB	1.86 (1.53-2.25)	2.49e-10	1.43e-11
Consolidation AI	IGR	1.87 (1.26-2.77)	0.00196	
Disease extent AI	KB	2.14 (1.77-2.6)	7.11e-15	3.13e-16
Disease extent AI	IGR	1.87 (1.28-2.73)	0.00109	

Supp Table 2: association of lesion volumes inferred by *AI-segment* and severity.

AUC (O2>=15L/min or ventilation or death)				
Model description	**Additional variables**	**KB**	**IGR**	**CV KB**
Radiologist report		0.718 (0.625 - 0.811)	0.658 (0.559 - 0.756)	0.699 (0.563 - 0.813)
AI-severity		0.756 (0.665 - 0.847)	0.704 (0.611 - 0.797)	0.765 (0.631 - 0.854)
AI-segment		0.681 (0.580 - 0.783)	0.681 (0.581 - 0.780)	0.708 (0.581 - 0.822)
Clinical and bio		0.719 (0.630 - 0.807)	0.746 (0.648 - 0.844)	0.789 (0.692 - 0.875)
Trimodal radiologist report	10 best clinical and bio	0.782 (0.701 - 0.862)	0.748 (0.655 - 0.842)	0.795 (0.708 - 0.901)
Trimodal AI-severity	10 best clinical and bio	0.780 (0.700 - 0.860)	0.752 (0.656 - 0.849)	0.819 (0.707 - 0.896)
Trimodal AI-segment	10 best clinical and bio	0.756 (0.672 - 0.839)	0.761 (0.670 - 0.852)	0.808 (0.684 - 0.893)
AUC DEATH				
Radiologist report		0.655 (0.543 - 0.768)	0.638 (0.518 - 0.759)	0.637 (0.476 - 0.791)
AI-severity		0.807 (0.724 - 0.890)	0.729 (0.626 - 0.833)	0.730 (0.610 - 0.853)
AI-segment		0.639 (0.522 - 0.756)	0.685 (0.558 - 0.811)	0.644 (0.499 - 0.795)
Clinical and bio		0.792 (0.695 - 0.890)	0.872 (0.797 - 0.948)	0.817 (0.709 - 0.931)
Trimodal radiologist report	10 best clinical and bio	0.849 (0.771 - 0.926)	0.856 (0.771 - 0.942)	0.821 (0.711 - 0.937)
Trimodal AI-severity	10 best clinical and bio	0.827 (0.747 - 0.908)	0.886 (0.816 - 0.957)	0.822 (0.727 - 0.938)
Trimodal AI-segment	10 best clinical and bio	0.816 (0.737 - 0.895)	0.867 (0.790 - 0.944)	0.816 (0.735 - 0.946)
MIT COVID Analytics		0.853 (0.786 - 0.920)	0.650 (0.540 - 0.760)	
AUC DEATH OR ICU				
Radiologist report		0.728 (0.643 - 0.812)	0.741 (0.651 - 0.830)	0.716 (0.608 - 0.822)
AI-severity		0.765 (0.683 - 0.847)	0.800 (0.722 - 0.879)	0.759 (0.662 - 0.865)
AI-segment		0.685 (0.591 - 0.778)	0.762 (0.672 - 0.852)	0.714 (0.608 - 0.833)
Clinical and bio		0.767 (0.684 - 0.850)	0.803 (0.722 - 0.884)	0.800 (0.683 - 0.877)
Trimodal radiologist report	10 best clinical and bio	0.808 (0.731 - 0.885)	0.826 (0.746 - 0.907)	0.795 (0.691 - 0.882)
Trimodal AI-severity	10 best clinical and bio	0.806 (0.733 - 0.880)	0.835 (0.760 - 0.910)	0.819 (0.713 - 0.890)
Trimodal AI-segment	10 best clinical and bio	0.792 (0.712 - 0.872)	0.846 (0.774 - 0.918)	0.811 (0.707 - 0.896)
Colombi et al.		0.684 (0.594 - 0.774)	0.617 (0.518 - 0.716)	

Supp Table 3: AUC values for the different models on the different sets. Each model was trained on 646 patients from KB. Results are reported on the validation set from KB (150 patients) and the external validation set from IGR (137 patients), as well as on the training set using 5 fold cross validation stratified by outcome and age (CV KB).

References of the Supplementary Material

1. 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).
2. Hansell, D. M. *et al.* Fleischner Society: glossary of terms for thoracic imaging. *Radiology* **246**, 697–722 (2008).
3. 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).
4. 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).
5. 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).
6. 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).
7. Hofmanninger, J. *et al.* Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem. *arXiv [eess.IV]* (2020).
8. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv [cs.LG]* (2019).

9. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv [cs.CV]* (2015).
10. Chen, X., Fan, H., Girshick, R. & He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv [cs.CV]* (2020).
11. 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).
12. LIDC-IDRI - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.
13. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
14. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).
15. Colombi, D. *et al.* Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology* 201433 (2020).