Radiology

Automated Assessment of CO-RADS and Chest CT Severity Scores in Patients with Suspected COVID-19 Using Artificial Intelligence

Nikolas Lessmann, PhD[1]; Clara I. Sánchez, PhD[1*]; Ludo Beenen, MD, PhD[2]; Luuk H. Boulogne, MSc[1]; Monique Brink, MD, PhD[1]; Erdi Calli, MSc[1]; Jean-Paul Charbonnier, PhD[3]; Ton Dofferhoff, MD, PhD[12]; Wouter M. van Everdingen, MD, PhD[1]; Paul K. Gerke, MSc[1]; Bram Geurts, MD[1]; Hester A. Gietema, MD, PhD[4,5]; Miriam Groeneveld, MSc[1]; Louis van Harten, MSc[6]; Nils Hendrix, MA[1]; Ward Hendrix, MSc[1]; Henkjan J. Huisman, PhD[1]; Ivana Išgum, PhD[6]; Colin Jacobs, PhD[1]; Ruben Kluge, BSc[1]; Michel Kok, MSc[1]; Jasenko Krdzalic, MD[7]; Bianca Lassen-Schmidt, PhD[8]; Kicky van Leeuwen, MSc[1]; James Meakin, DPhil[1]; Mike Overkamp, BSc[1]; Tjalco van Rees Vellinga, MD[9]; Eva M. van Rikxoort, PhD[3]; Riccardo Samperna, MSc[1]; Cornelia Schaefer-Prokop, MD, PhD[10,1]; Steven Schalekamp, MD, PhD[1,10]; Ernst Th. Scholten, MD, PhD[1]; Cheryl Sital, MSc[1]; Lauran Stöger, MD, PhD[11]; Jonas Teuwen, PhD[1]; Kiran Vaidhya Venkadesh, MSc[1]; Coen de Vente, MSc[1]; Marieke Vermaat, MD[13]; Weiyi Xie, MSc[1]; Bram de Wilde, MSc[1]; Mathias Prokop, MD, PhD[1]; Bram van Ginneken, PhD[1]*

[1] Department of Radiology, Nuclear Medicine and Anatomy, Radboud University Medical Center, Nijmegen, The Netherlands.

[2] Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands

[3] Thirona, Nijmegen, the Netherlands

[4] Department of Radiology and Nuclear Medicine, Maastricht University Medical Center+, The Netherlands

[5] GROW School of Oncology and Developmental Biology, Maastricht, The Netherlands

[6] Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, Amsterdam, The Netherlands

[7] Department of Radiology, Zuyderland MC, Heerlen, The Netherlands

[8] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

[9] Department of Radiology and Nuclear Medicine, Haaglanden Medical Center, The Hague, The Netherlands

[10] Department of Radiology, Meander Medical Center, Amersfoort, The Netherlands

[11] Department of Radiology, Leiden University Medical Centre, Leiden, The Netherlands

[12] Department of Internal Medicine, Canisius-Wilhelmina Ziekenhuis, Nijmegen, The Netherlands

[13] Department of Radiology, Canisius-Wilhelmina Ziekenhuis, Nijmegen, The Netherlands

* N.L and C.I.S contributed equally to this work

Address for correspondence:

Bram van Ginneken, PhD

Department of Radiology, Nuclear Medicine and Anatomy, Radboud University Medical Center

P.O. Box 9101

6500 HB Nijmegen, The Netherlands

E-mail: bram.vanginneken@radboudumc.nl

Word count: 2640 (without references and figure legends), abstract: 239

Article Type: Original Research

Summary statement: CORADS-AI is a freely accessible deep learning algorithm that automatically assigns CO-RADS and CT severity scores to non-contrast CT scans of patients suspected of COVID-19 with high diagnostic performance.

Key Results:

- CORADS-AI predicted scores from chest CT exams that were within one CO-RADS category in 81% of the patients and within one CT severity score point per lobe in 94% of the patients compared to readings by eight independent human observers.
- CORADS-AI identified patients with COVID-19 from chest CT exams with an AUC of 0.95 in an internal cohort and with an AUC of 0.88 in an external cohort.

Abbreviations: AI: Artificial Intelligence, AUC: Area under the Receiver Operating Characteristic curve, CI: confidence interval; CO-RADS: COVID-19 Reporting and Data System; CTSS: CT severity score; RT-PCR: Reverse Transcription Polymerase Chain Reaction, ROC: Receiver Operating Characteristic.

Abstract

Background: The COVID-19 pandemic has spread across the globe with alarming speed, morbidity and mortality. Immediate triage of suspected patients with chest infections caused by COVID-19 using chest CT may be of assistance when results from definitive viral testing are delayed.

Purpose: To develop and validate an artificial intelligence (AI) system to score the likelihood and extent of pulmonary COVID-19 on chest CT scans using the CO-RADS and CT severity scoring systems.

Materials and Methods: CORADS-AI consists of three deep learning algorithms that automatically segment the five pulmonary lobes, assign a CO-RADS score for the suspicion of COVID-19 and assign a CT severity score for the degree of parenchymal involvement per lobe. This study retrospectively included patients who received an unenhanced chest CT scan due to clinical suspicion of COVID-19 at two medical centers. The system was trained, validated, and tested with data from one of the centers. Data from the second center served as an external test set. Diagnostic performance and agreement with scores assigned by eight independent observers were measured using receiver operating characteristic (ROC) analysis, linearly-weighted kappa and classification accuracy.

Results: 105 patients (62 ± 16 years, 61 men) and 262 patients (64 ± 16 years, 154 men) were evaluated in the internal and the external test set, respectively. The system discriminated between COVID-19 positive and negative patients with areas under the ROC curve of 0.95 (95% CI: 0.91-0.98) and 0.88 (95% CI: 0.84-0.93). Agreement with the eight human observers was moderate to substantial with a mean linearly-weighted kappa of 0.60 ± 0.01 for CO-RADS scores and 0.54 ± 0.01 for CT severity scores.

Conclusion: CORADS-AI correctly identified COVID-19 positive patients with high diagnostic performance from chest CT exams, assigned standardized CO-RADS and CT severity scores in good agreement with eight independent observers and generalized well to external data.

Introduction

During the COVID-19 pandemic, chest CT imaging has been found useful for diagnosis and follow-up of COVID-19 patients (1). Standardized CT scoring systems, such as CO-RADS (2), have been advocated to improve the communication between radiologists and other healthcare providers by translating radiological findings into standardized scores (2-4). The CO-RADS scoring system assigns scores from 1 to 5 that increase with the level of suspicion of COVID-19 based on features seen on an unenhanced chest CT. Additionally, beyond assessing the likelihood of COVID-19, this and similar scoring systems also report on the extent of parenchymal involvement by assigning a CT severity score (CTSS) to patients with high suspicion of COVID-19 (5, 6). Such standardized scoring systems enable fast and consistent clinical decision making, especially valuable in these enduring times of crisis (2, 3).

Artificial Intelligence (AI) using deep learning has been advocated for automated reading of COVID-19 CT scans, including diagnosing COVID-19 (7-14) and quantifying parenchymal involvement (15-18). While these studies illustrate the potential of AI algorithms, their practical value is debatable (19). Without adhering to radiological reporting standards, it is doubtful whether these algorithms provide any real benefit in addition to or instead of manual reading, limiting their adoption in daily practice. Also, algorithms that follow a standardized scoring system need validation to confirm that they assign scores in a similar way to radiologists and can identify COVID-19 positive patients with similar or even better performance.

The purpose of this study was to develop and validate an AI algorithm (CORADS-AI) that automatically scores chest CT scans of patients with suspected COVID-19 according to the CO-RADS and CT severity score systems. We compared CORADS-AI with readings of eight observers and to clinical assessment of the patients including RT-PCR testing.

Materials and Methods

Medical ethics committee approval was obtained prior to the study. The need for written informed consent was waived and data was collected and anonymized in accordance with local guidelines.

Study sample

We retrospectively included consecutive patients presenting at the emergency wards of an academic center and a large teaching hospital in the Netherlands in March and April 2020 and underwent chest CT imaging for clinical suspicion of moderate to severe COVID-19. Criteria for CT were symptoms of lower respiratory tract infection including cough, clinically relevant dyspnea requiring hospital admission and fever with anosmia. CO-RADS and CT severity scores were reported as part of routine interpretation of the scans. Patients without scores in their radiological report were excluded. Additionally, patients from the teaching hospital were excluded if they were known to have COVID-19 (RT-PCR proven) prior to imaging or if RT-PCR test results were missing. The CT scanners were from different manufacturers; the protocols are described in Appendix E1.

Since RT-PCR testing may initially produce false negative results, we considered patients positive for COVID-19 if they either had a positive RT-PCR or if their clinical presentation made COVID-19 probable. Criteria were the lack of an alternative diagnosis explaining the symptoms plus intensive care admission due to respiratory failure, the need for high oxygen delivery, or unexplained death during admission.

Training and development set

The data of 476 patients from the academic center was used for model development, 520 CT scans in total. CO-RADS scores were extracted from the radiological reports. Scans scored CO-RADS 6, which signifies a positive RT-PCR prior to imaging, were rescored

independently by a chest radiologist with more than 30 years of experience (E.Th.S.) who assigned CO-RADS 1-5 to simulate unknown RT-PCR status (n=52). This observer was blinded to the original radiological report and all non-imaging data except for age and sex.

Internal test set

A prior observer study assessing CO-RADS (2) reported on the remaining 105 patients included at the academic center. The data of these patients was set aside to verify the performance of the AI model. For each patient at least one RT-PCR was available within five days after CT acquisition. The earliest available scan of each patient was scored independently following the CO-RADS classification by seven chest radiologists and one radiology resident (B.G., J.K., L.F.B., M.P., H.A.G., J.L.S., C.M.S-P., T.R.V.) using a dedicated browser-based workstation (CIRRUS[1]). Any available later scans from these 105 patients were not used in this study. The observers were familiar with the CO-RADS and CT severity scoring systems from interpreting at least 30 scans. Four of them had <5 years of experience in reading chest CTs, the others up to 27 years. All observers were blinded to the RT-PCR test results and could therefore not assign CO-RADS 6. Instead, they assigned CO-RADS 1-5 based on their suspicion for pulmonary involvement of COVID-19. Also, they semi-quantitatively described the extent of parenchymal involvement per lobe using a predefined CT severity score on a 6-point scale (0=0%, 1=0%-5%, 2=5%-25%, 3=25%-50%, 4=50%-75%, 5=>75%) (5).

External test set

The data of all patients included at the teaching hospital was set aside to verify the performance of the AI model on an external cohort. All these patients underwent RT-PCR testing on the same day as CT imaging. The CO-RADS score and the total CT severity

[1] Available at *https://grand-challenge.org/reader-studies/*

score (the sum of the scores per lobe as described above) were extracted from the radiological report of the earliest available scan of each patient.

Annotation of pulmonary lobes and opacities

Reference delineations of lung and lobar boundaries were automatically obtained for a convenience sample of 400 scans from the training and development set and for all 105 scans in the internal test set using commercial software (LungQ v1.1.1, Thirona, Nijmegen, The Netherlands), followed by manual correction. Reference delineations of areas with ground-glass opacities, consolidation, and mixed patterns were obtained for a convenience sample of 108 scans from the training and development set as follows. Regions of parenchymal lung tissue with increased attenuation were identified with thresholding and morphological operations. Vessels and airways were removed using automatic methods. Lesion candidates in lobes not affected by COVID-19 following the radiological report were removed. The remaining lesion candidates were reviewed by a certified image analyst with at least one year of experience in correcting segmentations of pulmonary structures on chest CT exams. The analyst corrected the delineations and added and removed lesions as needed.

Automated CT scoring

CT scans were scored fully automatically using three successively applied deep learning algorithms. These performed (1) pulmonary lobe segmentation and labeling, (2) lesion segmentation and CT severity score prediction, and (3) CO-RADS score prediction.

For lobe segmentation and labeling, we used a relational two-stage U-Net architecture specifically developed for robust pulmonary lobe segmentation (20). The model was pre-trained on 4000 chest CT scans from the COPDGene study (21) and fine-tuned with 400 scans from the present study.

For CT severity score prediction, we trained a 3D U-net using the nnU-Net framework (22) in a cross-validated fashion with 108 scans and corresponding reference delineations to segment ground-glass opacities and consolidation in the lungs. The CT severity score was derived from the segmentation results by computing the percentage of affected parenchymal tissue per lobe.

For CO-RADS score prediction, we used the 3D-inflated Inception architecture (23, 24), a 3D extension of the state-of-the-art Inception image classification architecture (25). The model was pre-trained on the ImageNet (26) and Kinetics (27) data sets and trained with 368 CT scans from our present study to predict the corresponding CO-RADS score. The remaining scans from the training and development set were used to monitor the performance during training. Input to the model was the CT image together with areas of abnormal parenchymal lung tissue detected by the severity scoring algorithm.

Further details about the methods are provided in Appendix E2. The algorithm is freely accessible online[2].

Statistical analysis

Lobe segmentation was evaluated using the average Dice coefficient per lobe in the internal test set. Diagnostic performance of the automated CO-RADS scoring algorithm was evaluated using receiver operating characteristic curves (ROC) and the area under the ROC curve (AUC). Youden's index was used to determine the optimal threshold. Non-parametric bootstrapping with 1000 iterations was used to calculate 95% confidence intervals (95% CI). To quantify agreement, linearly-weighted kappa and classification accuracy were determined by comparing the predicted CO-RADS and CT severity scores to the median of all

[2] https://grand-challenge.org/algorithms/corads-ai/

combinations of seven observers. The agreement of the AI system in terms of the linearly-weighted kappa was compared with the agreement of the left-out observer using Monte-Carlo permutation tests. CT severity scores were evaluated only for patients with a diagnosis of COVID-19. We tested for differences in demographic characteristics between training and test cohorts using t-tests (age) and chi-squared tests (sex), and for differences in sensitivity of observers and algorithm at the specificity of the observers using McNemar tests. The significance level was .05. Analyses were performed with statistical software (R 3.6.2) and Python 3.7.6 (scipy 1.5.0, sklearn 0.23.1, evalutils 0.2.3).

Results

Patient characteristics

A total of 581 and 262 consecutive patients were included at the academic center and the teaching hospital, respectively. Training and development set comprised 520 scans of 476 patients from the academic center, the internal test set comprised 105 scans of the remaining 105 patients. The external test set comprised 262 scans of 262 patients from the teaching hospital. Six patients were excluded due to CO-RADS scores missing from their radiological reports (Figure 1). Patient characteristics for training and test sets are given in Table 1.

There were 58 patients with a clinical diagnosis of COVID-19 among the 105 patients in the internal test set (55%) and 179 among the 262 patients in the external test set (68%). Of these patients, 53/58 (91%) and 145/179 (81%) had a positive RT-PCR while the remaining positive patients had one or multiple negative RT-PCR tests but were diagnosed with COVID-19 based on their clinical presentation. Table 2 summarizes the distribution of CO-RADS and CT severity scores according to the radiological reports. The algorithm was executed successfully on all scans in the test sets. For the 105 scans in the internal test set,

median runtime of the algorithm was 212 seconds (range 146-709), while the median reading time by radiologists was 82 seconds (range 58-134).

Lobe segmentation

Reference delineations of lung and lobar boundaries were available for 104 of the 105 scans in the internal test set. In one image, the lobar boundaries could not be identified due to severe emphysema and the presence of an open window thoracostomy after a Clagett procedure in the right lung. In the remaining 104 images, the average Dice scores of the automatic lobe segmentations were 95.2% ± 2.0% for the left upper lobes, 92.4% ± 10.1% for the left lower lobes, 95.2% ± 3.1% for the right upper lobes, 92.2% ± 10.7% for the right middle lobes and 94.7% ± 3.7% for the right lower lobes.

Identification of COVID-19 patients

In the internal test set, the algorithm distinguished between negative and positive COVID-19 patients with an AUC of 0.95 (95% CI: 0.91-0.98) based on the CO-RADS 5 probability that the algorithm predicted**.** At the optimal threshold, the sensitivity of the algorithm was 85.7% (95% CI: 73.1-98.2) and the specificity was 89.8% (95% CI: 79.6-100). In the external test set, the AUC was 0.88 (95% CI: 0.84-0.93) and sensitivity and specificity at the optimal threshold were 82% (95% CI: 69.7-94.3) and 80.5% (95% CI: 67.9-93.1), respectively. The corresponding ROC curves are shown in Figure 2 together with the operating points of the eight observers for the internal test set, and with the operating points for the routinely reported CO-RADS scores for the external test set. In the internal test set, the mean (± standard deviation) sensitivity of the eight observers was 61.4% (± 7.9%) at a mean specificity of 99.7% (± 0.7%) based on patients that they scored CO-RADS 5. In the external test set, the CO-RADS scores reported as part of clinical routine corresponded to a sensitivity of 134 of 179 (74.9%) and a specificity of 74 of 83 (89.2%). The sensitivities and specificities of the observers for each operating point and the sensitivities of the AI algorithm

at the same specificities are given in Table 3. There was enough evidence for significantly better sensitivity of the observer for only 3 of the tested 36 operating points (8.3%) of all observers combined.

CO-RADS score prediction

When compared with the median CO-RADS score of all combinations of seven of the eight readers of the internal test set, the automatically assigned CO-RADS score was in absolute agreement in 54.8% (460 of 8x105=840) of the patients and within one category in 80.5% (676/840). The remaining reader was in absolute agreement in 68.2% (573/840) of the patients and within one category in 96.2% (808/840). In the external test set, AI algorithm and reference score were in absolute agreement in 64.12% (168/262) and within one category in 85.50% (224/262) of the patients. The cross-tabulated results are given in Appendix E3.

There was moderate to substantial agreement between AI-predicted CO-RADS scores and the observers according to the linearly-weighted kappa (Table 4). For the internal test set, the mean (± standard deviation) kappa was 0.60 ± 0.01 for the AI system and 0.79 ± 0.04 for the left-out observer ($p<0.001$ for all observers). For the external test set, kappa was 0.69 (95% CI: 0.63-0.75) for the AI system.

CT severity score prediction

Since the automatic prediction is based on a segmentation of the lobes and abnormal regions in the lung, the algorithm outputs the percentage of affected parenchymal tissue rather than just the categorical severity score. Figure 3 depicts the percentage of affected parenchymal tissue per lobe with respect to the median severity score of the readers in the internal test set. The predicted score was in absolute agreement with the median score of all combinations of seven of the eight observers in 17.2% (80 of 8x58 positive patients = 464)

and deviated by not more than one point per lobe, i.e., 5 points in total, in 94.0% (436/464). In the external test set, the radiological reports contained severity scores for 163 of the 179 positive patients. The AI algorithm was in absolute agreement with these scores in 17 of 163 (10.4%) patients and within one point per lobe in 146 of 163 (89.6%).

There was moderate agreement between AI-predicted severity scores and the observers according to the linearly-weighted kappa (Table 4). For the internal test set, the mean (± standard deviation) kappa was 0.54 ± 0.01 for the AI system and 0.77 ± 0.03 for the left-out observer ($p<0.001$ for all observers). For the external test set, kappa was 0.49 (95% CI: 0.41-0.56) for the AI system.

Representative examples of lobe segmentation results, CO-RADS score predictions and CT severity score predictions with corresponding pulmonary lesions are shown in Figures 4 to 6 and in Appendix E4.

Discussion

AI might be helpful in interpreting CT scans of patients with high suspicion of COVID-19, especially when it produces standardized output that radiologists and other healthcare providers are familiar with. In this study, we evaluated the performance of an AI system for automated scoring of chest CT scans of patients suspected of COVID-19 based on the CO-RADS and CT severity score classifications. This system identified COVID-19 patients with high diagnostic performance, achieving an AUC of 0.95 on the test set with similar sensitivity and specificity as the eight observers, and 0.88 on the external test set. The automated CO-RADS was in good agreement with the observers, although agreement was significantly higher between observers (mean kappa of 0.60 for AI vs. 0.79 inter-observer, $p<0.001$). Likewise, the automated CT severity score agreed well with the observers with a mean kapa of 0.54 ± 0.01, which was also lower than the agreement between the observers (kappa of

0.77 ± 0.03, p<0.001). An explanation may be that visually estimating the amount of affected lung parenchyma is subjective; studies have shown that human readers tend to overestimate the extent of disease (28). In the four cases where automatic measurements were >10 points higher than the reference, underlying causes were severe motion artifacts in three cases and one patient with opacifications caused by aspiration pneumonia. This underlines the importance of verification of automatically determined severity scores by humans.

In the short period since the outbreak, many groups have already developed AI algorithms for diagnosing COVID-19. Shi et al.(29) and Ito et al.(30) provide overviews of the proposed approaches. Most studies analyze small data sets and employ two-dimensional (2D) neural network architectures on axial sections. One of the first large studies by Li et al. (9) with such a 2D approach showed performance on an independent test set comparable to ours but did not compare AI results with human reading. We experimented with their CovNet architecture but found that our three-dimensional inflated Inception approach gave higher performance. Zhang et al. (17) also followed a 2D approach and showed on a large data set that segmentation of lesions per axial section and feeding these slices into a classification network gave excellent results on several Chinese data sets, outperforming junior radiologists. Their method separates consolidations and ground glass lesions, and this seems a promising approach. We followed the CT severity score that is part of CO-RADS and therefore have not separated different types of lesions. When tested on data from Ecuador, performance of their system dropped considerably. We also saw reduced performance on the external validation set for our method, but for that set the CO-RADS scores obtained from routine practice also showed lower performance, comparable to the AI system. Similarly, Bai et al. (12) obtained good results with a 2D approach (AUC 0.95), reducing to an AUC 0.90 on external data. To optimize AI software for data from different hospitals, Xu et al.(31) implemented a federated learning approach for COVID-19 diagnosis from CT and showed this may lead to better performance on external data.

Mei et al. (18) developed a COVID-19 AI diagnosis system and tested it on a held-out set of 279 patients. They followed a similar 2D approach as the methods discussed previously, but they added a separate neural network analyzing clinical symptoms and five laboratory findings. The combination of both networks provided the best performance (AUC 0.92). Given the overlap of morphological features with other non-COVID-19-related diseases, our study could be advanced with an AI analysis of imaging, clinical signs, and routine laboratory parameters. There are other studies that used visual scoring of imaging data and have demonstrated the potential of such a combined approach (32, 33).

None of the previously published works followed a standardized reporting scheme, and therefore these systems provide an unstandardized, uncalibrated output. The lack of explainability of the output of AI systems is often seen as a limitation. By adhering to a standardized reporting system already validated in human observers (2), the proposed system overcomes this obstacle. The lobar output of the CT severity system is also familiar to radiologists. We did not find any other publications where the accuracy of automated lobar segmentation and lesion segmentation in CT scans of COVID-19 suspects was quantitatively evaluated and compared with human readers.

Our study has several limitations. First, we trained the AI system with data from a single medical center. More training is needed with multi-center data sets. Second, the study sample represented the population of patients that presented in a high-prevalence situation at our hospital and receive chest CT scans due to suspicion of COVID-19. There were only a limited number of patients with extensive pre-existing lung disease in the training set. Third, inclusion of the study group took place after the influenza and RS-virus season. Consequently, most CT images in this study were either normal or demonstrated COVID-19 features. Ultimately, AI systems need to be trained with a larger dataset before they can be expected to correctly interpret studies with overlapping abnormalities due to other types of

pneumonia or other diseases such as congestive heart failure, pulmonary fibrosis or acute respiratory distress syndrome in non-COVID-19 patients.

In conclusion, our study demonstrates that an AI system can identify COVID-19 positive patients based on unenhanced chest CT images with diagnostic performance comparable to that of radiological observers. It is noteworthy that the algorithm was trained to adhere to the CO-RADS categories and is thus directly interpretable by radiologists. We believe the AI system may be of use to support radiologists in standardized CT reporting during busy periods. The automatically assessed CT severity scores per lobe may have prognostic information and could be used to quantify lung damage, also during patient follow-up.

References

1. Yang W, Sirajuddin A, Zhang X, Liu G, Teng Z, Zhao S, Lu M. The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). European Radiology 2020. doi: 10.1007/s00330-020-06827-4

2. Prokop M, van Everdingen W, van Rees Vellinga T, Quarles van Ufford J, Stöger L, Beenen L, Geurts B, Gietema H, Krdzalic J, Schaefer-Prokop C, van Ginneken B, Brink M. CO-RADS – A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation. Radiology 2020. doi: 10.1148/radiol.2020201473

3. Simpson S, Kay FU, Abbara S, Bhalla S, Chung JH, Chung M, Henry TS, Kanne JP, Kligerman S, Ko JP, Litt H. 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 2020;2(2). doi: 10.1148/ryct.2020200152

4. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus disease 2019 (COVID-19) imaging reporting and data system (COVID-RADS) and common lexicon: a proposal based on the imaging data of 37 studies. European Radiology 2020. doi: 10.1007/s00330-020-06863-0

5. Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, Li C. The Clinical and Chest CT Features Associated With Severe and Critical COVID-19 Pneumonia. Invest Radiol 2020;55(6):327-331. doi: 10.1097/RLI.0000000000000672

6. Chang YC, Yu CJ, Chang SC, Galvin JR, Liu HM, Hsiao CH, Kuo PH, Chen KY, Franks TJ, Huang KM, Yang PC. Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT. Radiology 2005;236(3):1067-1075. doi: 10.1148/radiol.2363040958

7. Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia. Applied Intelligence 2020. doi: 10.1007/s10489-020-01714-3

8. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, Chen J, Zhao H, Jie Y, Wang R, Chong Y, Shen J, Zha Y, Yang Y. Deep learning Enables Accurate Diagnosis of Novel

Coronavirus (COVID-19) with CT images. medRxiv 2020:2020.2002.2023.20026930. doi: 10.1101/2020.02.23.20026930

9. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang G, Xu Q, Fang X, Zhang S, Xia J, Xia J. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology 2020. doi: 10.1148/radiol.2020200905

10. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, Cai M, Yang J, Li Y, Meng X, Xu B. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv 2020:2020.2002.2014.20023028. doi: 10.1101/2020.02.14.20023028

11. Jin S, Wang B, Xu H, Luo C, Wei L, Zhao W, Hou X, Ma W, Xu Z, Zheng Z, Sun W, Lan L, Zhang W, Mu X, Shi C, Wang Z, Lee J, Jin Z, Lin M, Jin H, Zhang L, Guo J, Zhao B, Ren Z, Wang S, You Z, Dong J, Wang X, Wang J, Xu W. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks. medRxiv 2020:2020.2003.2019.20039354. doi: 10.1101/2020.03.19.20039354

12. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, Tran TML, Choi JW, Wang D-C, Shi L-B, Mei J, Jiang X-L, Pan I, Zeng Q-H, Hu P-F, Li Y-H, Fu F-X, Huang RY, Sebro R, Yu Q-Z, Atalay MK, Liao W-H. AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT. Radiology 2020. doi: 10.1148/radiol.2020201491

13. Ouyang X, Huo J, Xia L, Shan F, Liu J, Mo Z, Yan F, Ding Z, Yang Q, Song B, Shi F, Yuan H, Wei Y, Cao X, Gao Y, Wu D, Wang Q, Shen D. Dual-Sampling Attention Network for Diagnosis of COVID-19 from Community Acquired Pneumonia. IEEE Transactions on Medical Imaging 2020:1-1. doi: 10.1109/tmi.2020.2995508

14. Wang J, Bao Y, Wen Y, Lu H, Luo H, Xiang Y, Li X, Liu C, Qian D. Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images. IEEE Transactions on Medical Imaging 2020:1-1. doi: 10.1109/tmi.2020.2994908

15. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, Bernheim A, Siegel E. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for

Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. arXiv e-prints 2020;arXiv:2003.05037. Accessed March 01, 2020.

16. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shen D, Shi Y. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. arXiv e-prints 2020;arXiv:2003.04655. Accessed March 01, 2020.

17. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, Liang W, Wang C, Wang K, Ye L, Gao M, Zhou Z, Li L, Wang J, Yang Z, Cai H, Xu J, Yang L, Cai W, Xu W, Wu S, Zhang W, Jiang S, Zheng L, Zhang X, Wang L, Lu L, Li J, Yin H, Wang W, Li O, Zhang C, Liang L, Wu T, Deng R, Wei K, Zhou Y, Chen T, Lau JY-N, Fok M, He J, Lin T, Li W, Wang G. 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

18. Mei X, Lee H-C, Diao K-y, Huang M, Lin B, Liu C, Xie Z, Ma Y, Robson PM, Chung M, Bernheim A, Mani V, Calcagno C, Li K, Li S, Shan H, Lv J, Zhao T, Xia J, Long Q, Steinberger S, Jacobi A, Deyer T, Luksza M, Liu F, Little BP, Fayad ZA, Yang Y. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Medicine 2020. doi: 10.1038/s41591-020-0931-3

19. Kundu S, Elhalawani H, Gichoya JW, Kahn CE. How Might AI and Chest Imaging Help Unravel COVID-19's Mysteries? Radiology: Artificial Intelligence 2020;2(3). doi: 10.1148/ryai.2020200053

20. Xie W, Jacobs C, Charbonnier J-P, van Ginneken B. Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans. IEEE Transactions on Medical Imaging 2020:1-1. doi: 10.1109/tmi.2020.2995108

21. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD: Journal of Chronic Obstructive Pulmonary Disease 2010;7(1):32-43. doi: 10.3109/15412550903499522

22. Isensee F, Jäger PF, Kohl SAA, Petersen J, Maier-Hein KH. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. arXiv e-prints 2019;arXiv:1904.08128. Accessed April 01, 2019.

23. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2017; p. 4724-4733.

24. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP, Shetty S. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 2019;25(6):954-961. doi: 10.1038/s41591-019-0447-x

25. Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2015; p. 1-9.

26. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 2015;115(3):211-252. doi: 10.1007/s11263-015-0816-y

27. Li A, Thotakuri M, Ross DA, Carreira J, Vostrikov A, Zisserman A. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv e-prints 2020;arXiv:2005.00214. Accessed May 01, 2020.

28. Gietema HA, Muller NL, Fauerbach PV, Sharma S, Edwards LD, Camp PG, Coxson HO, Evaluation of CLtIPSEi. Quantifying the extent of emphysema: factors associated with radiologists' estimations and quantitative indices of emphysema severity using the ECLIPSE cohort. Acad Radiol 2011;18(6):661-671. doi: 10.1016/j.acra.2011.01.011

29. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, He K, Shi Y, Shen D. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Rev Biomed Eng 2020. doi: 10.1109/RBME.2020.2987975

30. Ito R, Iwano S, Naganawa S. A review on the use of artificial intelligence for medical imaging of the lungs of patients with coronavirus disease 2019. Diagn Interv Radiol 2020. doi: 10.5152/dir.2019.20294

31. Xu Y, Ma L, Yang F, Chen Y, Ma K, Yang J, Yang X, Chen Y, Shu C, Fan Z, Gan J, Zou X, Huang R, Zhang C, Liu X, Tu D, Xu C, Zhang W, Yang D, Wang MW, Wang X, Xie X, Leng H, Holalkere N, Halin NJ, Kamel IR, Wu J, Peng X, Wang X, Shao J, Mongkolwat P, Zhang J, Rubin DL, Wang G, Zheng C, Li Z, Bai X, Xia T. A collaborative online AI engine for CT-based COVID-19 diagnosis. medRxiv 2020. doi: 10.1101/2020.05.10.20096073

32. Dofferhoff ASM, Swinkels A, Sprong T, Berk Y, Spanbroek M, Nabuurs-Franssen MH, Vermaat M, van de Kerkhof B, Willekens MHC, Voss A. [Diagnostic algorithm for COVID-19 at the ER]. Ned Tijdschr Geneeskd 2020;164.

33. Kurstjens S, van der Horst A, Herpers R, Geerits MWL, Kluiters-de Hingh YCM, Gottgens EL, Blaauw MJT, Thelen MHM, Elisen M, Kusters R. Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing. Clin Chem Lab Med 2020. doi: 10.1515/cclm-2020-0593

Table 1. Characteristics of the Training and Test Cohorts

	Training set (n=476)	Internal test set (n=105)		External test set (n = 262)	
Age (mean ± SD)	60 ± 16	62 ± 16	p=0.20	64 ± 16	p<0.01
Sex			p=0.79		p=0.68
Men	267 (56)	61 (58)		154 (54)	
Women	209 (44)	44 (42)		120 (46)	
Number of CT scans per patient					
1	438 (92)	105 (100)		262 (100)	
2	34 (7)	-		-	
3	2 (0)	-		-	
4	2 (0)	-		-	

Data are numbers of patients with percentages in parentheses unless otherwise noted.
P values correspond to comparisons of the distribution of characteristics in the training set and the respective test set.

Table 2. CO-RADS and CT Severity Scores according to the Radiological Reports or according to the Radiologist Who Reviewed Scans Initially Scored CO-RADS 6 (proven COVID-19)

	Training set (n=520)	Internal test set (n=105)	External test set (n=262)
CO-RADS categories			
CO-RADS 1	236 (45)	20 (19)	56 (21)
CO-RADS 2	66 (13)	10 (10)	12 (5)
CO-RADS 3	80 (15)	19 (18)	26 (10)
CO-RADS 4	45 (9)	17 (16)	25 (10)
CO-RADS 5	93 (18)	39 (37)	143 (54)
CT Severity Score	8 (9)	10 (7)	11 (7)

CO-RADS data are numbers of CT scans with percentages in parentheses. CT severity score data are median and interquartile ranges and were reported only for cases with CO-RADS ≥ 3.

Table 3. Sensitivity of the Observers and the AI Algorithm for Identification of COVID-19 at the Specificity Levels Corresponding to Various Operating Points of the Observers

	Specificity	Observer sensitivity	AI sensitivity	p-value
Internal test set (n=105)				
CO-RADS 5				
Observer 1	100%	60.3%	71.9% (95% CI: 58.7-84.8)	0.143
Observer 2	97.9%	74.1%	73.0% (95% CI: 59.3-88.3)	>0.999
Observer 3	100%	62.1%	71.9% (95% CI: 58.7-84.8)	0.238
Observer 4	100%	46.6%	71.9% (95% CI: 58.7-84.8)	<0.001
Observer 5	100%	55.2%	71.9% (95% CI: 58.7-84.8)	0.064
Observer 6	100%	58.6%	71.9% (95% CI: 58.7-84.8)	0.143
Observer 7	100%	69%	71.9% (95% CI: 58.7-84.8)	>0.999
Observer 8	100%	65.5%	71.9% (95% CI: 58.7-84.8)	0.607
CO-RADS 4+5				
Observer 1	95.7%	87.9%	76.1% (95% CI: 61.2-92.3)	0.146
Observer 2	95.7%	86.2%	76.1% (95% CI: 61.2-92.3)	0.302
Observer 3	97.9%	77.6%	73.0% (95% CI: 59.3-88.3)	0.791
Observer 4	97.9%	77.6%	73.0% (95% CI: 59.3-88.3)	0.791
Observer 5	100%	72.4%	71.9% (95% CI: 58.7-84.8)	>0.999
Observer 6	100%	84.5%	71.9% (95% CI: 58.7-84.8)	0.118
Observer 7	97.9%	86.2%	73.0% (95% CI: 59.3-88.3)	0.180
Observer 8	100%	75.9%	71.9% (95% CI: 58.7-84.8)	0.815
CO-RADS 3+4+5				
Observer 1	63.8%	98.3%	97.2% (95% CI: 90.9-100.0)	>0.999
Observer 2	61.7%	98.3%	97.5% (95% CI: 91.4-100.0)	>0.999
Observer 3	76.6%	93.1%	93.6% (95% CI: 83.1-100.0)	>0.999
Observer 4	74.5%	91.4%	94.5% (95% CI: 84.7-100.0)	0.549
Observer 5	89.4%	93.1%	85.6% (95% CI: 69.2-96.6)	0.607

Observer 6	85.1%	93.1%	89.0% (95% CI: 75.8-98.2)	0.791
Observer 7	76.6%	94.8%	93.6% (95% CI: 83.1-100.0)	>0.999
Observer 8	87.2%	87.3%	87.3% (95% CI: 72.0-96.9)	0.815
CO-RADS ≥ 2				
Observer 1	23.4%	100%	99.1% (95% CI: 94.7-100.0)	>0.999
Observer 2	34%	100%	98.2% (95% CI: 93.8-100.0)	>0.999
Observer 3	38.3%	98.3%	98.2% (95% CI: 93.8-100.0)	>0.999
Observer 4	14.9%	98.3%	99.9% (95% CI: 98.1-100.0)	0.754
Observer 5	53.2%	98.3%	98.1% (95% CI: 93.1-100.0)	>0.999
Observer 6	53.2%	98.3%	98.1% (95% CI: 93.1-100.0)	>0.999
Observer 7	55.3%	98.3%	98.0% (95% CI: 92.9-100.0)	>0.999
Observer 8	51.1%	93.1%	98.1% (95% CI: 93.1-100.0)	0.503
External test set (n=262)				
CO-RADS 5	89.2%	74.9%	67.6% (95% CI: 54.6-81.7)	0.035
CO-RADS 4+5	83.1%	86%	77.1% (95% CI: 63.1-88.6)	0.020
CO-RADS 3+4+5	62.7%	91.1%	92.5% (95% CI: 86.3-96.7)	0.585
CO-RADS ≥ 2	54.2%	93.9%	94.2% (95% CI: 90.1-97.7)	>0.999

P values correspond to comparisons of the sensitivity of the observer with that of the AI algorithm.

Table 4. Agreement of the Observers and the AI System with the Median Score Assigned by the Remaining Seven Observers in the Internal Test Set

Left-out observer	Left-out observer vs. median observer	AI algorithm vs. median observer	p-value
CO-RADS classification			
Observer 1	0.79 (0.73-0.86)	0.59 (0.50-0.70)	<0.001
Observer 2	0.78 (0.71-0.86)	0.59 (0.49-0.70)	<0.001
Observer 3	0.85 (0.80-0.90)	0.60 (0.50-0.69)	<0.001
Observer 4	0.72 (0.65-0.79)	0.60 (0.50-0.71)	0.018
Observer 5	0.74 (0.66-0.81)	0.61 (0.51-0.71)	0.032
Observer 6	0.84 (0.79-0.90)	0.61 (0.52-0.71)	<0.001
Observer 7	0.82 (0.77-0.88)	0.58 (0.48-0.68)	<0.001
Observer 8	0.79 (0.73-0.86)	0.61 (0.51-0.70)	0.002
Total CT severity score			
Observer 1	0.80 (0.74-0.85)	0.53 (0.44-0.63)	<0.001
Observer 2	0.74 (0.68-0.81)	0.53 (0.43-0.62)	<0.001
Observer 3	0.78 (0.71-0.84)	0.54 (0.45-0.63)	<0.001
Observer 4	0.76 (0.69-0.84)	0.53 (0.44-0.63)	<0.001
Observer 5	0.80 (0.74-0.87)	0.54 (0.45-0.63)	<0.001
Observer 6	0.71 (0.64-0.79)	0.55 (0.45-0.64)	0.004
Observer 7	0.81 (0.75-0.87)	0.53 (0.44-0.63)	<0.001
Observer 8	0.73 (0.65-0.80)	0.55 (0.46-0.64)	0.002

Data are linearly-weighted kappa scores with 95% confidence intervals in parentheses. *P* values correspond to comparisons of the kappa scores of left-out observer and AI system using the median of the other seven observers as reference.

FIGURE LEGENDS

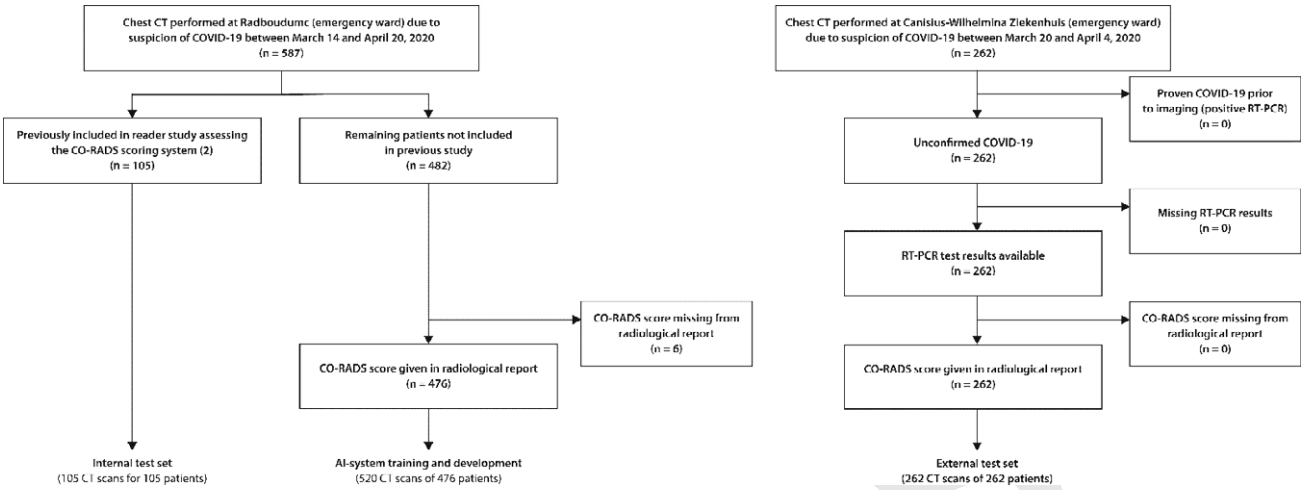

Figure 1. Flowchart for patient inclusion in the training and test sets.

Note that n refers to the number of patients. The number of CT images is higher in the training set since several patients received multiple chest CT scans within the inclusion period. However, in the test sets, only the earliest available scan of each patient was used.

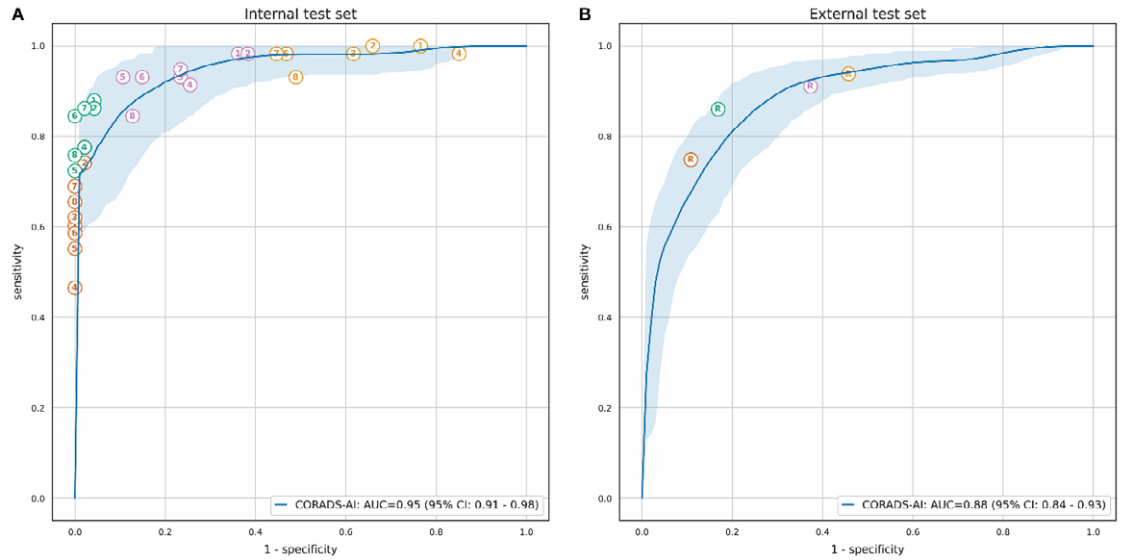

Figure 2. ROC curves for automatically predicted CO-RADS 5 probability vs. COVID-19 diagnosis.

The receiver operating characteristic (ROC) curve is based on the probability that the algorithm assigned to CO-RADS 5. The shaded area around the ROC curve reflects the 95% confidence interval. *A,* The performance of the eight observers is shown as individual points on the graph for the internal test set, and, *B,* the diagnostic performance of the scores from the radiological reports is shown for the external test set. Different colors indicate different cut-offs, where patients were considered predicted COVID-19 positive if the observer assigned a CO-RADS score of 5 (orange), 4 or 5 (green), 3 to 5 (magenta), or 2 to 5 (yellow). COVID-19 diagnosis meant either a positive RT-PCR test or very high clinical suspicion of COVID-19 despite at least one negative RT-PCR test.

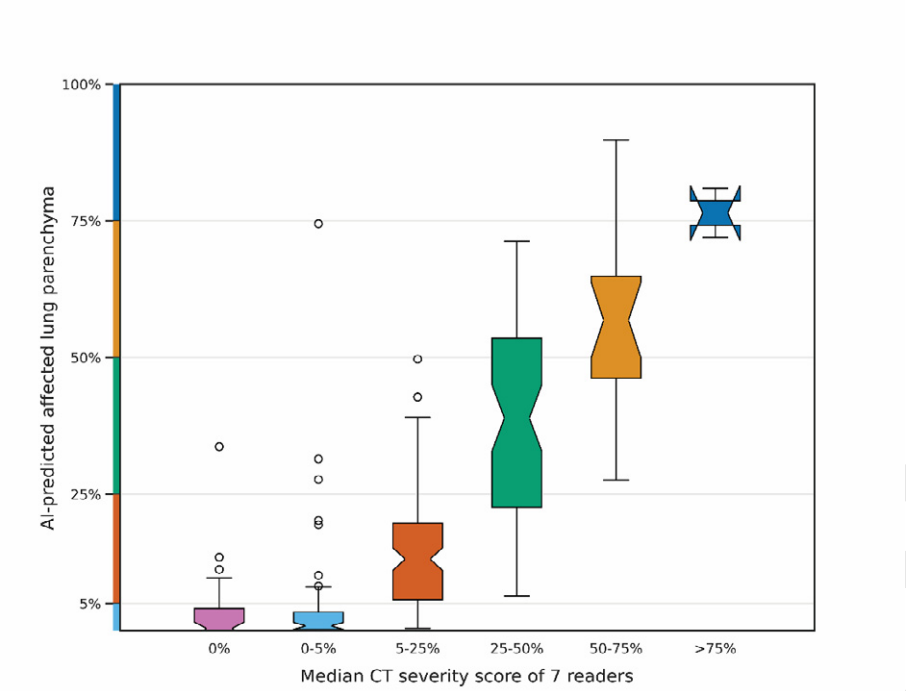

Figure 3. CT severity score predictions vs. median of observer scores.

Shown as box plots are the distribution of the percentage of affected lung parenchyma per lobe according to the automatic lesion (affected volume) and lobe segmentations (total volume) for the internal test set. The notch in each box plot illustrates the 95% confidence interval around the median. The CT severity score cut-offs are marked on the y-axis.

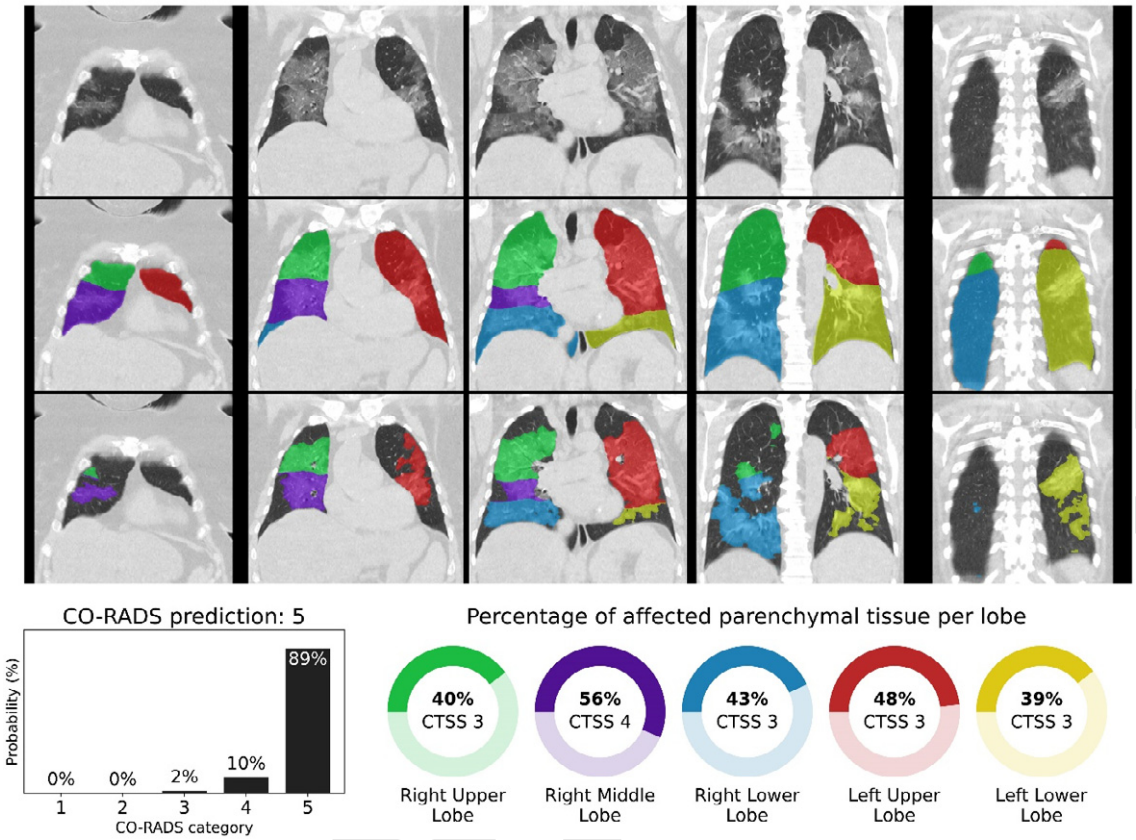

Figure 4. CO-RADS and CT severity score predictions for a COVID-19 positive case with extensive parenchymal involvement.

73-year-old woman with positive RT-PCR test result. Non-contrast CT scan in coronal view (top row), overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row). This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), and the computed percentages of affected parenchymal tissue and the corresponding CT Severity Score (CTSS) per lobe per lobe (bottom right). The eight observers scored this case 3x CO-RADS 3, 1x CO-RADS 4 and 4x CO-RADS 5.

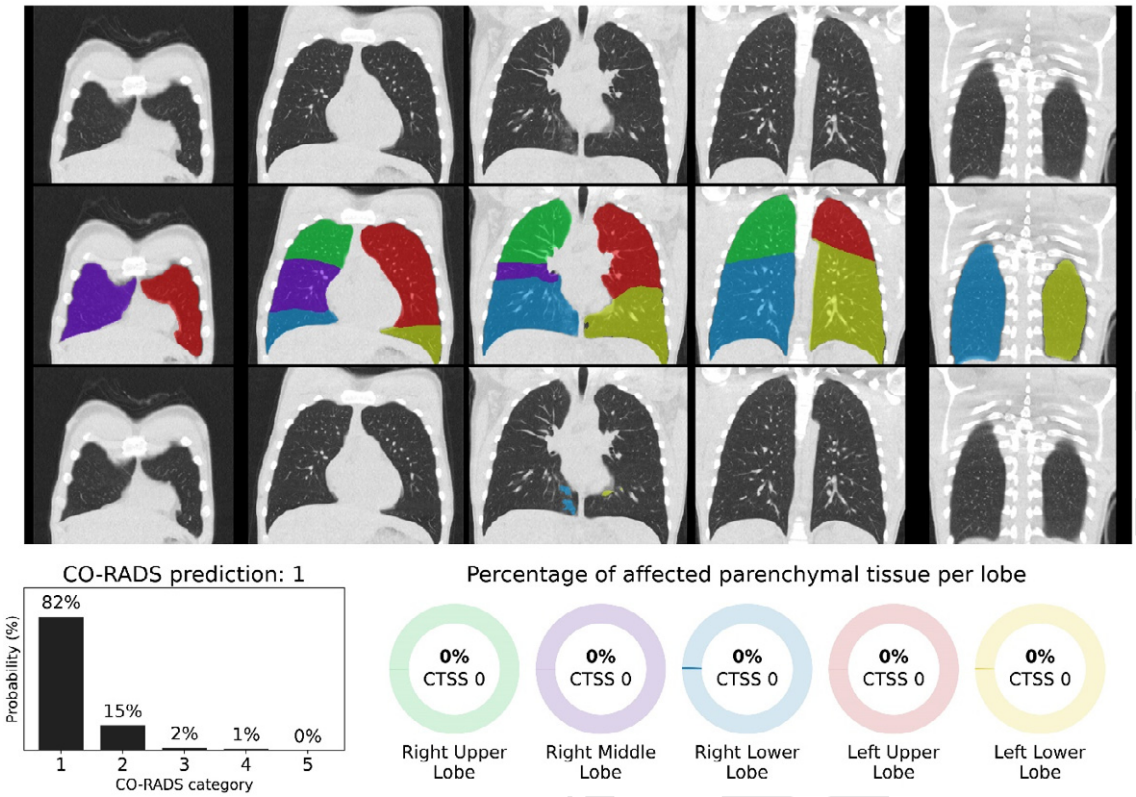

Figure 5. CO-RADS and CT severity score predictions for a COVID-19 positive case with little parenchymal involvement.

18-year-old man with positive RT-PCR test result. Non-contrast CT scan in coronal view (top row), overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row). This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), and the computed percentages of affected parenchymal tissue and the corresponding CT Severity Score (CTSS) per lobe (bottom right). The eight observers scored this case 2x CO-RADS 1, 5x CO-RADS 2 and 1x CO-RADS 3.

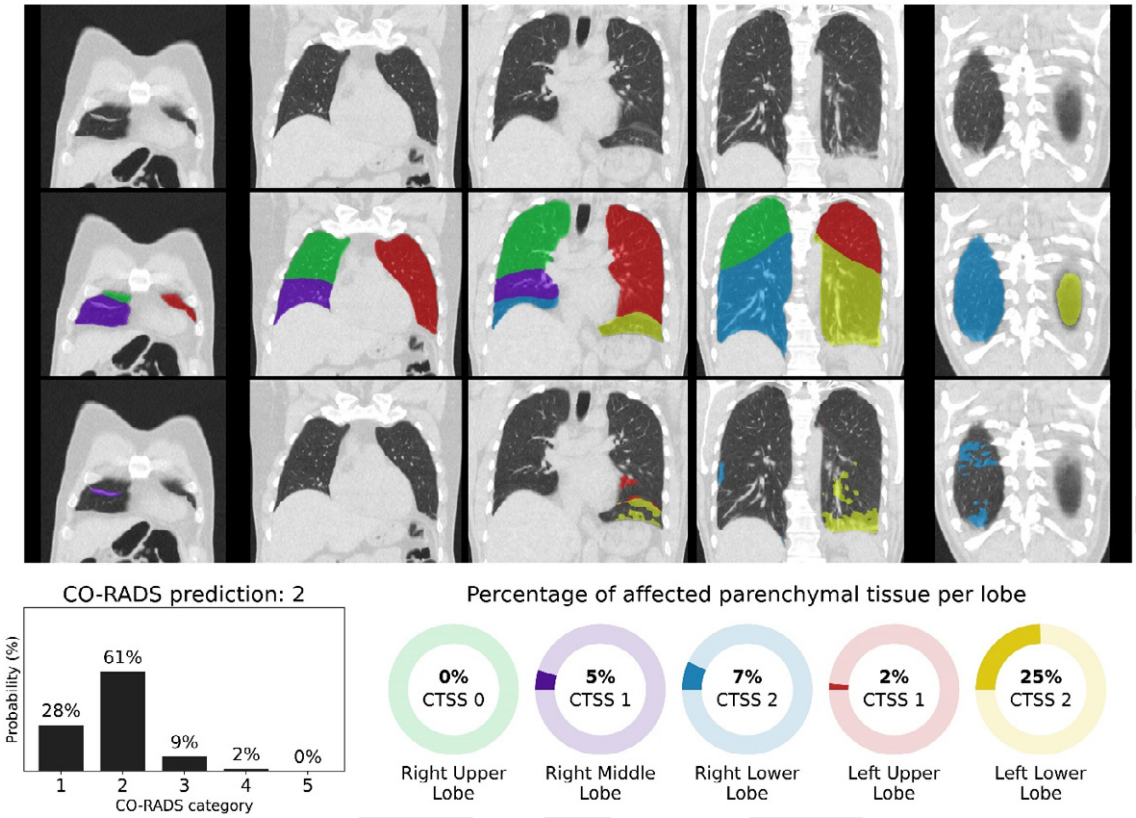

Figure 6. CO-RADS and CT severity score predictions for a COVID-19 negative case. 54-year-old man with negative RT-PCR test result. Non-contrast CT scan in coronal view (top row), overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row). This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), and the computed percentages of affected parenchymal tissue and the corresponding CT Severity Score (CTSS) per lobe (bottom right). The eight observers scored this case 3x CO-RADS 1, 3x CO-RADS 2, and 2x CO-RADS 3.

Appendix E1: CT Protocol

At the academic center, scans were obtained on 160 or 320 multidetector CT (Canon Medical Systems, Otawara, Japan) without intravenous contrast injection, with a tube energy of 100, 120 or 135kV depending on body weight, collimation of 80 x 0.5mm, rotation time of 0.275s, and were reconstructed using hybrid iterative reconstruction (AIDR 3D) with a pre-set noise tolerance of 35 into axial slices with 0.5mm thickness and 0.5mm increment. Median $CTDI_{vol}$ was 0.9mGy (IQR 0.6 – 1.6), and median DLP was 30.9 mGy*cm (IQR 18.2 – 50.3).

At the teaching hospital, scans were obtained on a 128 multidetector CT (Philips Medical Systems, Best, Netherlands) without intravenous contrast injection, with a tube energy of 100kV, collimation of 64 x 0.625mm, rotation time of 0.4s, and were reconstructed using model-based iterative reconstruction (IMR1, SharpPlus kernel) into axial slices with 1mm thickness and 1mm increment. Median $CTDI_{vol}$ was 2.9mGy (IQR 1.8 – 4.2), and median DLP was 109.5 mGy*cm (IQR 72.3 – 162.0).

In both centers, a deep inspiration breath-hold technique was applied whenever feasible.

Appendix E2: Components of the AI System

The CORADS-AI system for automatic assessment of CO-RADS score and CT Severity Score in chest CT scans of patients with suspected COVID-19 is composed of three successively applied deep learning algorithms for (1) pulmonary lobe segmentation and labeling, (2) CT severity score prediction and (3) CO-RADS score prediction. Following standard practice, we chose the architectures and hyper-parameters that obtained the highest accuracy on our development set (520 CT images). The test sets were not used for model development.

The first component of the CORADS-AI system automatically segments the pulmonary lobes in a CT scan using a deep learning algorithm based on a relational two-stage U-Net architecture (RTSU-Net) [1]. RTSU-Net is based on a cascade of two 3D U-Net architectures [2] and integrates a relational module to capture the global context of a large 3D CT volume using appearance and geometric information. This architecture takes a full CT volume as input and outputs the corresponding segmentation mask of the five pulmonary lobes. The algorithm was first trained with 4000 chest CT scans from the COPDGene study [3] and then fine-tuned with 400 scans from the present study to increase the robustness to COVID-19 typical pulmonary abnormalities, such as consolidated areas. The remaining scans were used for validation during training, i.e., for tuning the model parameters. Further details of the RTSU-Net architecture and training procedure can be found elsewhere [1].

Prior to executing the second and third components, which perform automated CO-RADS and CT severity scoring, the original CT volumes were resampled to an isotropic resolution of 1.5 mm x 1.5 mm x 1.5 mm voxel spacing to ensure that all images are interpreted at the same resolution. The volumes were cropped to the lungs based on the automatic lobe segmentation by cropping to the minimum bounding box of the union of all five lobes plus a 10 mm additional margin on each side. Since the CO-RADS model made use of transfer

learning and therefore expected input values in the range 0 to 1, the images were additionally normalized for this model by clipping intensities below -1100 HU and above 300 HU and mapping the range -1100 HU to 300 HU to the range 0 to 1.

For CT severity score prediction, we trained a 3D U-net using the nnU-Net framework[3] [4] with 108 scans in which we semi-automatically delineated ground-glass opacities and pulmonary consolidation. The network was trained using a five-fold cross-validation scheme. The resulting probability maps were resampled to the original resolution of the CT scan and then thresholded at 50% probability. False-positive detections outside the lobes were automatically removed using the lobe segmentation mask. The volume of all detected lesions per lobe was divided by the volume of the corresponding lobe and the resulting percentage of affected parenchymal tissue was used to derive the CT severity score (0% = CTSS 0, 0% - 5% = CTSS 1, 5% - 25% = CTSS 2, 25% - 50% = CTSS 3, 50% - 75% = CTSS 4, >75% = CTSS 5).

For CO-RADS score prediction, we used a deep learning model based on a 3D inflated Inception V1 architecture (I3D)[4] [5,6]. The I3D architecture builds upon the state-of-the-art Inception v1 model for 2D image classification but *inflates* the filters and pooling kernels into 3D. This enables the use of an image classification model pre-trained with 2D data for a 3D image classification task. The model was pre-trained in 2D with the ImageNet dataset [7], converted into a 3D model and then further pre-trained using the Kinetics dataset [8], a 3D image classification dataset. The resulting pre-trained 3D image classification model was trained with 368 scans of patients with a clinical suspicion of COVID-19 for CO-RADS score prediction, the remaining scans were used for validation during training, i.e., for tuning the model parameters. Because the pre-trained I3D models expect the 3D input data to have the same dimensions as the data the model was pre-trained with, the cropped and isotropically

[3] Source code available at *https://github.com/MIC-DKFZ/nnUNet*
[4] Source code available at *https://github.com/deepmind/kinetics-i3d*

resampled CT volumes were further cropped or padded to a fixed size of 240 x 240 pixels in-plane. In axial direction, 128 slices were uniformly sampled from the lung region. Lesion masks generated by the CT severity score model were pre-processed to the same shape and fed to the model together with the CT image. Since the I3D model expected a 3-channel input due to pre-training with RGB data, the first channel was the image, the second the lesion mask and the third was all zeros. All batch normalization layers in the model were fixed after pre-training. We trained the model with a batch size of 2, a learning rate of 0.0001 with Adam optimizer and cross entropy loss.

References

1. Xie W, Jacobs C, Charbonnier J-P, van Ginneken B. Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans. IEEE Transactions on Medical Imaging 2020:1-1. doi: 10.1109/tmi.2020.2995108

2. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 20162016; p. 424-432.

3. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD: Journal of Chronic Obstructive Pulmonary Disease 2010;7(1):32-43. doi: 10.3109/15412550903499522

4. Isensee F, Jäger PF, Kohl SAA, Petersen J, Maier-Hein KH. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. arXiv e-prints 2019;arXiv:1904.08128. Accessed April 01, 2019.

5. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2017; p. 4724-4733.

6. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP, Shetty S. End-to-end lung cancer screening with three-dimensional

deep learning on low-dose chest computed tomography. Nature Medicine 2019;25(6):954-961. doi: 10.1038/s41591-019-0447-x

7. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 2015;115(3):211-252. doi: 10.1007/s11263-015-0816-y

8. Li A, Thotakuri M, Ross DA, Carreira J, Vostrikov A, Zisserman A. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv e-prints 2020;arXiv:2005.00214. Accessed May 01, 2020.

Appendix E3: Additional Results

Table C1: Scores of the AI System Cross-Tabulated against the Median Scores of All Combinations of Seven Observers for the Internal Test Set

		Median of all combinations of seven of the eight observers				
		CO-RADS 1	**CO-RADS 2**	**CO-RADS 3**	**CO-RADS 4**	**CO-RADS 5**
AI	**CO-RADS 1**	64/840 (8%)	56/840 (7%)	0/840 (0%)	0/840 (0%)	0/840 (0%)
	CO-RADS 2	32/840 (4%)	64/840 (8%)	0/840 (0%)	0/840 (0%)	0/840 (0%)
	CO-RADS 3	24/840 (3%)	24/840 (3%)	24/840 (3%)	0/840 (0%)	16/840 (2%)
	CO-RADS 4	16/840 (2%)	16/840 (3%)	48/840 (6%)	24/840 (3%)	16/840 (2%)
	CO-RADS 5	0/840 (0%)	8/840 (1%)	72/840 (8%)	56/840 (7%)	280/840 (33%)

Data corresponds to comparison of the AI-predicted CO-RADS scores with median scores of all leave-one-reader-out combinations, i.e., 8 x 105 patients = 840 reference CO-RADS scores.

Table C2: Scores of the AI System Cross-Tabulated against the Scores from the Radiological Reports of the Scans in the External Test Set

		Radiological report				
		CO-RADS 1	**CO-RADS 2**	**CO-RADS 3**	**CO-RADS 4**	**CO-RADS 5**
AI	**CO-RADS 1**	37/262 (14%)	5/262 (2%)	7/262 (3%)	2/262 (1%)	0/262 (0%)
	CO-RADS 2	11/262 (4%)	2/262 (1%)	4/262 (2%)	1/262 (0%)	1/262 (0%)
	CO-RADS 3	7/262 (3%)	4/262 (2%)	8/262 (3%)	13/262 (5%)	14/262 (5%)
	CO-RADS 4	1/262 (0%)	0/262 (0%)	3/262 (1%)	5/262 (2%)	12/262 (5%)
	CO-RADS 5	0/262 (0%)	1/262 (0%)	4/262 (2%)	4/262 (2%)	116/262 (44%)

Appendix E4: Example Output of the CORADS AI System

We randomly selected two images for each CO-RADS score category (1 to 5) from the 105 CT scans in the internal test set, i.e., images that the CORADS-AI system was not trained with, using the CO-RADS scores from the radiological reports. The images, segmentation results and CO-RADS score and CTSS predictions are shown in Figures D-1 to D-5. Note that the CT severity score is intended for assessment of patients with clinical suspicion of COVID-19 who require hospitalization and is therefore typically not reported for unsuspicious cases. The radiological reports of these 10 example cases therefore contained CT severity scores only for cases scored CO-RADS 3 or higher, these are reported in the figure captions.

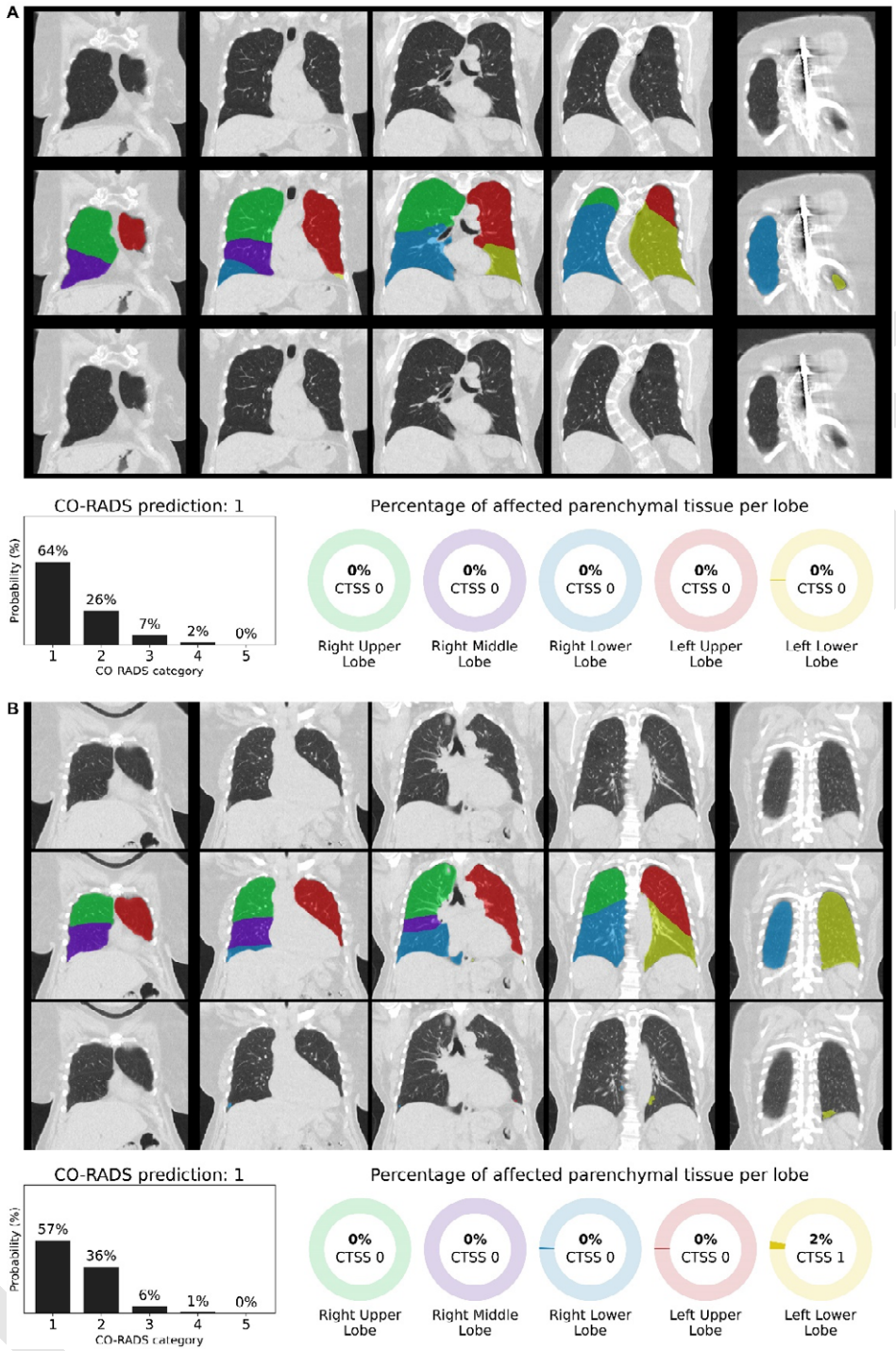

Figure D1. Model output for CO-RADS 1 cases according to the radiological report

Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 8x CO-RADS 1 (top), and 3x CO-RADS 1, 4x CO-RADS 2 and 1x CO-RADS 3 (bottom).

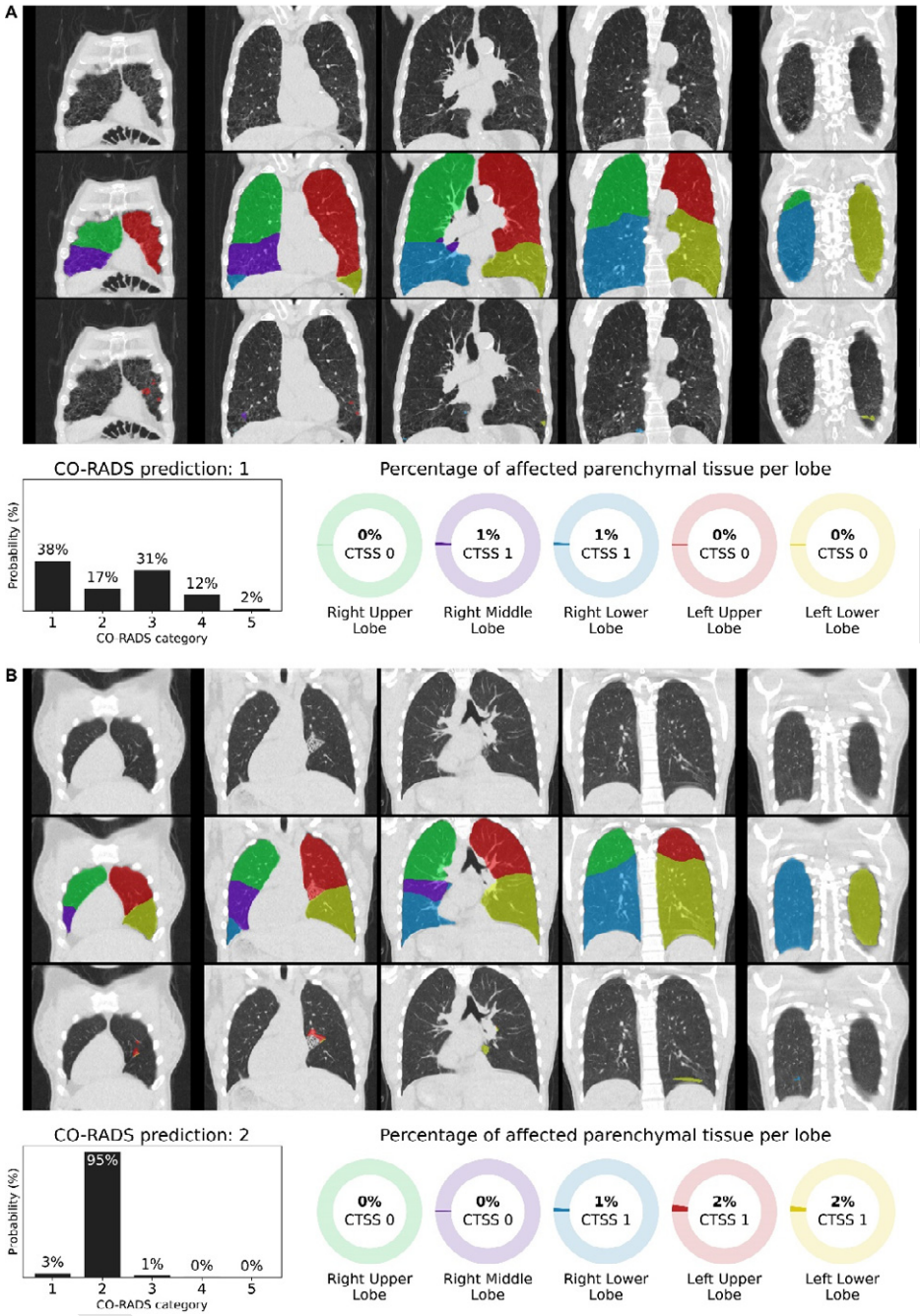

Figure D2. Model output for CO-RADS 2 cases according to the radiological report
Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3x CO-RADS 1 and 5x CO-RADS 2 (top), and 4x CO-RADS 1, 1x CO-RADS 2 and 3x CO-RADS 3 (bottom). There was evidence of extensive emphysema with bilateral areas of ground glass (top) and of situs inversus (bottom).

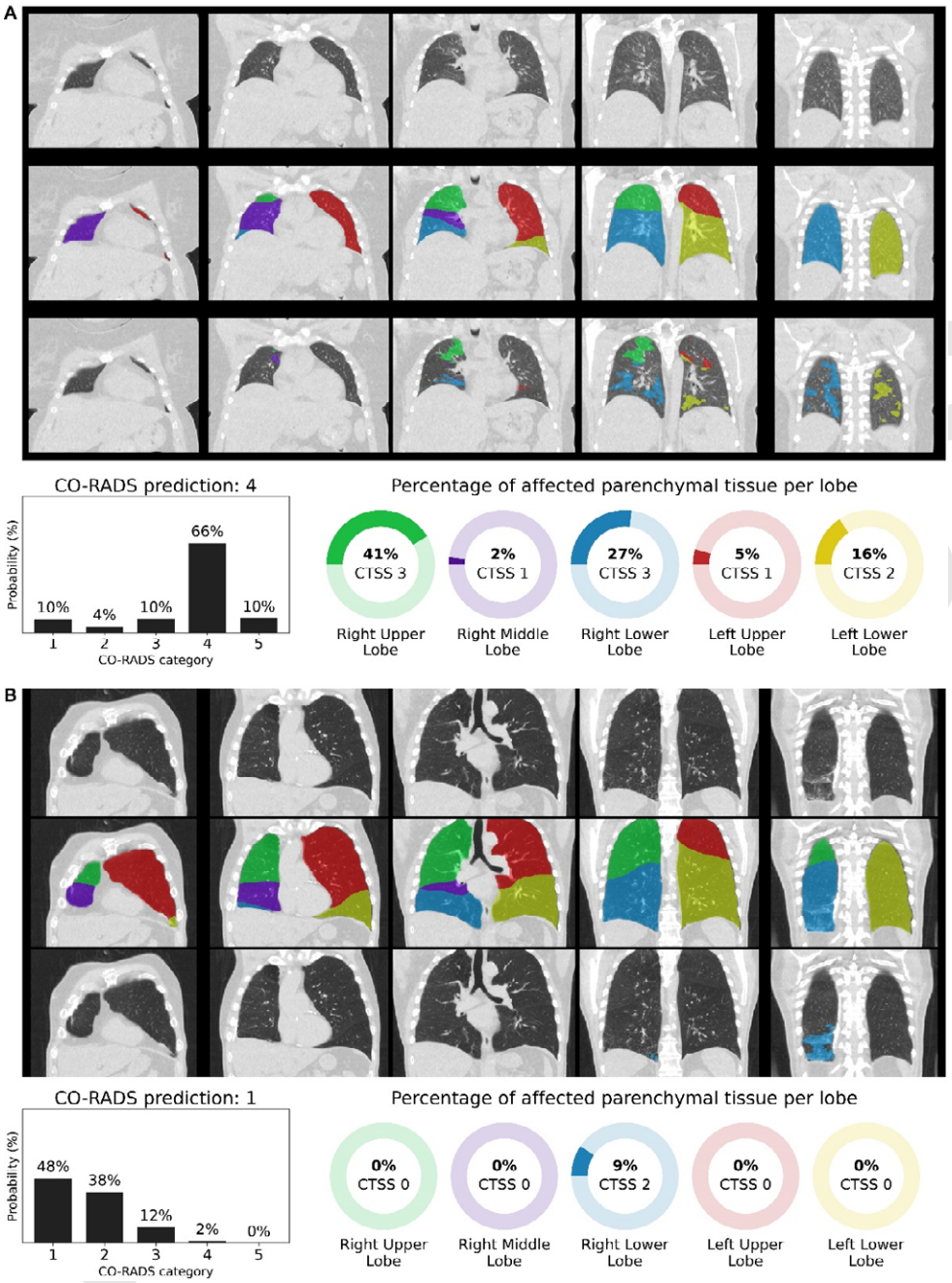

Figure D3. Model output for CO-RADS 3 cases with CTSS 12 (top) and CTSS 2 (bottom) according to the radiological report

Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3x CO-RADS 1, 2x CO-RADS 2 and 3x CO-RADS 3 (top), and 8x CO-RADS 2 (bottom). There was evidence of small ground glass areas in the peri-hilar regions (top) and patchy consolidation in the right lower lobe (bottom).

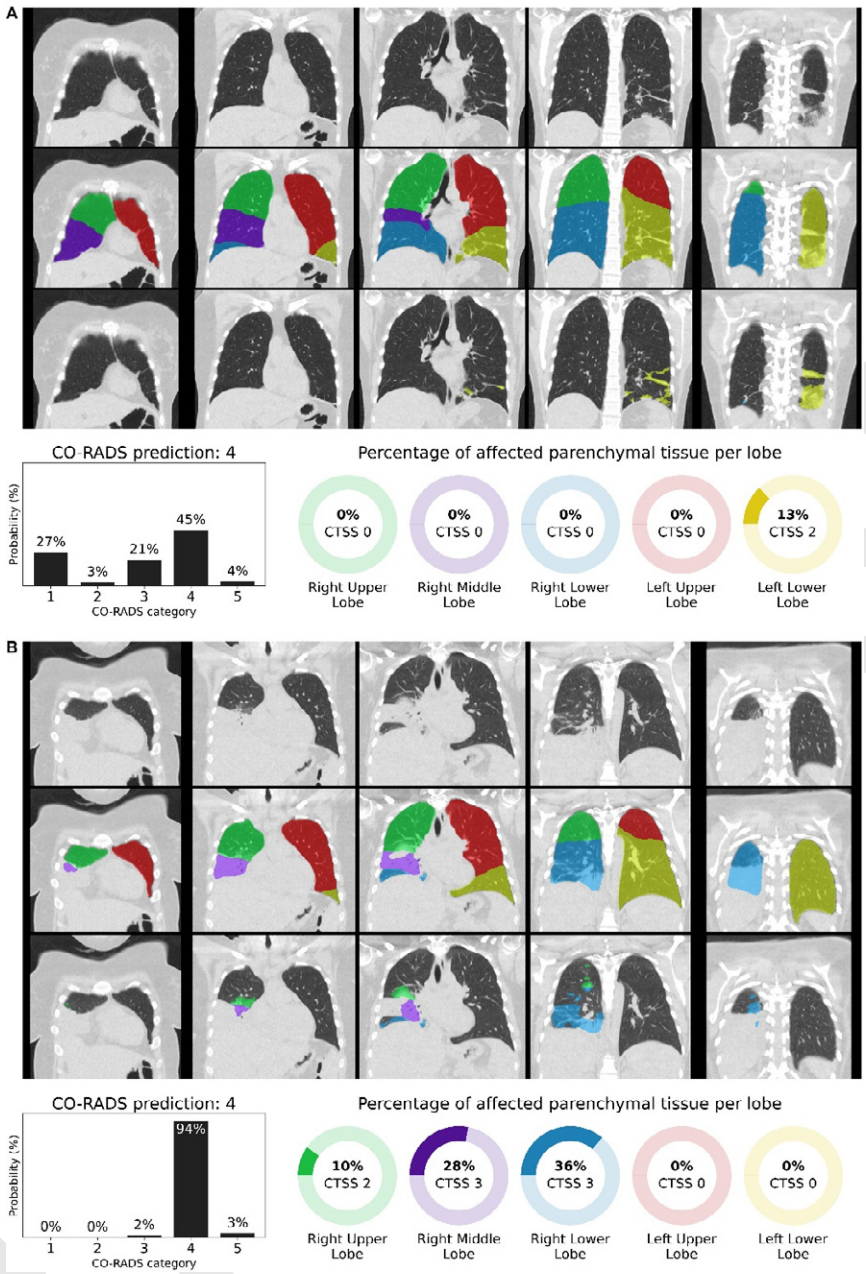

Figure D4. Model output for CO-RADS 4 cases with CTSS 5 (top) and CTSS 11 (bottom) according to the radiological report

Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3x CO-RADS 2, 4x CO-RADS 3 and 1x CO-RADS 4 (top), and 3x CO-RADS 3 and 5x CO-RADS 4 (bottom). There was evidence of ground glass areas in the periphery of the left lower lobe and areas of band-like collapse bilaterally (top), and of ground glass in the right upper lobe, dense consolidation in the entire middle lobe and posterior-basal in the right lower lobe (bottom).

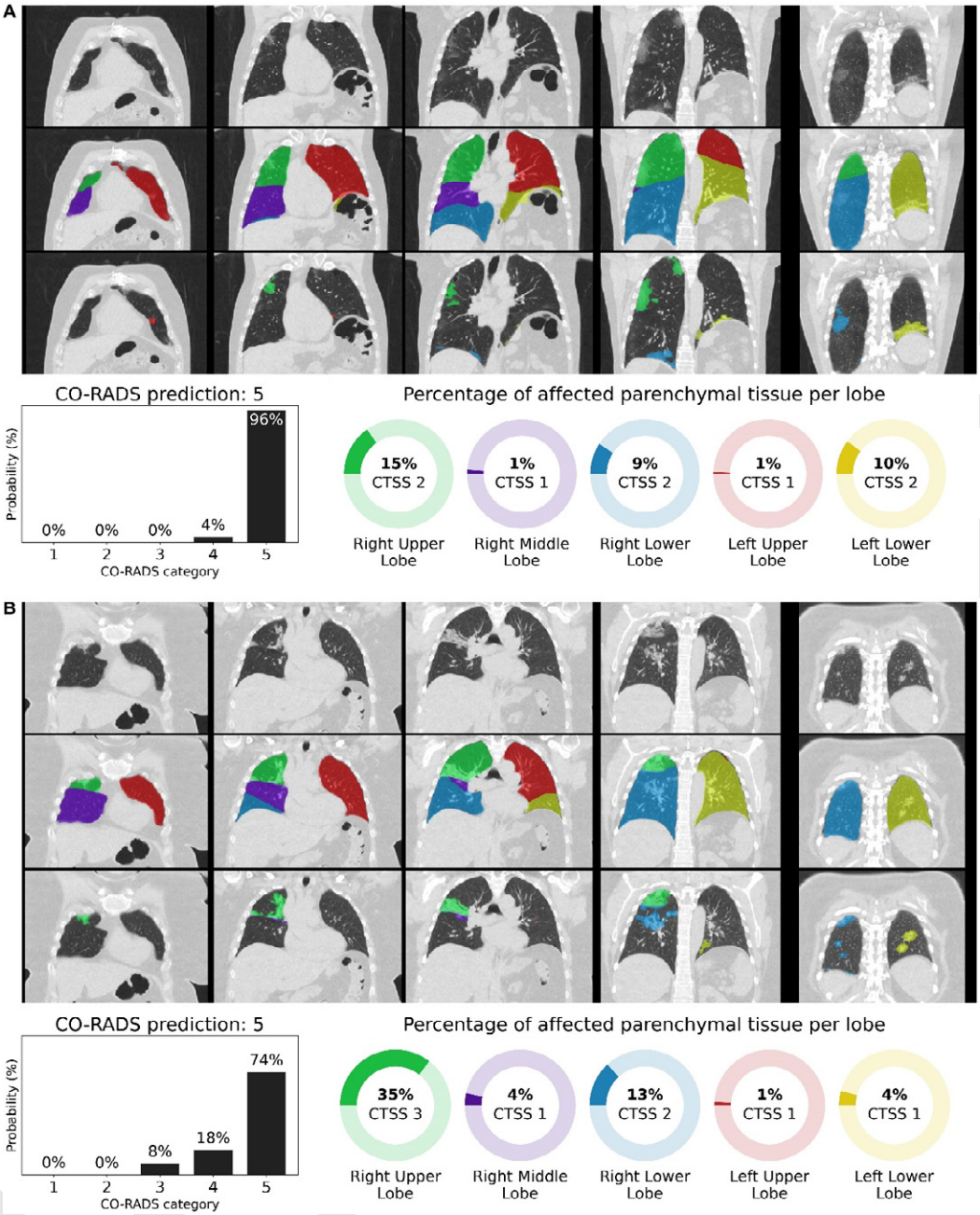

Figure D5. Model output for CO-RADS 5 cases with CTSS 8 (top) and CTSS 9 (bottom) according to the radiological report

Shown are coronal slices (top row) overlaid with the lobe segmentation results (middle row) and the lesion segmentation results and the corresponding CT Severity Score (CTSS) per lobe (bottom row). The eight observers scored the scans 3x CO-RADS 4 and 5x CO-RADS 5 (top), and 2x CO-RADS 4 and 6x CO-RADS 5 (bottom). There was evidence of areas of ground glass peripheral in the right lung, especially the upper lobe (top) and of extensive ground glass peripheral in the right lung as well as small areas in the left lung.