Radiology

Automated Assessment of COVID-19 Reporting and Data System and Chest CT Severity Scores in Patients Suspected of Having COVID-19 Using Artificial Intelligence

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

From the Department of Radiology, Nuclear Medicine and Anatomy, Radboud University Medical Center, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands (N.L., C.I.S., L.H.B., M.B., E.C., W.M.v.E., P.K.G., B.G., M.G., N.H., W.H., H.J.H., C.J., R.K., M.K., K.v.L., J.M., M.O., R.S., C. Schaefer-Prokop, S.S., E.T.S., C. Sital, J.T., K.V.V., C.d.V., W.X., B.d.W., M.P., B.v.G.); Department of Radiology, Academic Medical Center, Amsterdam, the Netherlands (L.B.); Thirona, Nijmegen, the Netherlands (J.P.C., E.M.v.R.); Departments of Internal Medicine (T.D.) and Radiology (M.V.), Canisius-Wilhelmina Ziekenhuis, Nijmegen, the Netherlands; Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht, the Netherlands (H.A.G.); GROW School of Oncology and Developmental Biology, Maastricht, the Netherlands (H.A.G.); Departments of Biomedical Physics and Engineering and Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, the Netherlands (L.v.H., I.I.); Department of Radiology, Zuyderland Medical Center, Heerlen, the Netherlands (J.K.); Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany (B.L.); Department of Radiology and Nuclear Medicine, Haaglanden Medical Center, The Hague, the Netherlands (T.v.R.V.); Department of Radiology, Meander Medical Center, Amersfoort, the Netherlands (C. Schaefer-Prokop, S.S.); and Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands (J.L.S.). Received May 27, 2020; revision requested June 22; revision received July 22; accepted July 29. **Address correspondence to** B.v.G. (e-mail: *bram.vanginneken@radboudumc.nl*).

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

Background: The coronavirus disease 2019 (COVID-19) pandemic has spread across the globe with alarming speed, morbidity, and mortality. Immediate triage of patients with chest infections suspected to be 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 COVID-19 Reporting and Data System (CO-RADS) and CT severity scoring systems.

Materials and Methods: The CO-RADS AI system 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 underwent a nonenhanced chest CT examination because of 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 analysis, linearly weighted κ values, and classification accuracy.

Results: A total of 105 patients (mean age, 62 years ± 16 [standard deviation]; 61 men) and 262 patients (mean age, 64 years ± 16; 154 men) were evaluated in the internal and external test sets, respectively. The system discriminated between patients with COVID-19 and those without COVID-19, with areas under the receiver operating characteristic curve of 0.95 (95% CI: 0.91, 0.98) and 0.88 (95% CI: 0.84, 0.93), for the internal and external test sets, respectively. Agreement with the eight human observers was moderate to substantial, with mean linearly weighted κ values of 0.60 ± 0.01 for CO-RADS scores and 0.54 ± 0.01 for CT severity scores.

Conclusion: With high diagnostic performance, the CO-RADS AI system correctly identified patients with COVID-19 using chest CT scans and assigned standardized CO-RADS and CT severity scores that demonstrated good agreement with findings from eight independent observers and generalized well to external data.

©RSNA, 2020

Supplemental material is available for this article.

During the coronavirus disease 2019 (COVID-19) pandemic, chest CT imaging has been found useful in the diagnosis and follow-up of patients with COVID-19 (1). Standardized CT scoring systems, such as the COVID-19 Reporting and Data System (CO-RADS) (2), have been advocated to improve communication between radiologists and other health care providers by translating radiologic findings into standardized scores (2–4). The CO-RADS

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019, RT-PCR = reverse transcription polymerase chain reaction

Summary

The Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) artificial intelligence system is a freely accessible deep learning algorithm that automatically assigns CO-RADS and CT severity scores to nonenhanced CT scans of patients suspected of having COVID-19 with high diagnostic performance.

Key Results

- The Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) artificial intelligence (AI) system assigned scores to CT scans that were within one CO-RADS category and within one per-lobe CT severity score point of the scores assigned by eight independent human observers in 81% and 94% of the patients evaluated, respectively.
- The CO-RADS AI system identified patients with COVID-19 using chest CT scans with an area under the receiver operating characteristic (AUC) curve of 0.95 in an internal cohort and an AUC of 0.88 in an external cohort.

scoring system assigns scores from 1 to 5 that increase with the level of suspicion of COVID-19 based on features seen on nonenhanced chest CT scans. 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 to patients highly suspected of having COVID-19 (5,6). Such standardized scoring systems enable fast and consistent clinical decision making, which is 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 those used to diagnose COVID-19 (7–14) and quantify parenchymal involvement (15–18). Although these studies illustrate the potential of AI algorithms, their practical value is debatable (19). Without adhering to radiologic reporting standards, it is doubtful that these algorithms provide any real benefit in addition to or instead of manual reading, limiting their adoption in daily practice. In addition, algorithms that follow a standardized scoring system need validation to confirm that they assign scores in a manner similar to that of radiologists and can be used to identify patients with COVID-19 with similar or even better performance.

The purpose of this study was to develop and validate an AI algorithm (the CO-RADS AI system) that automatically scores chest CT scans of patients suspected of having COVID-19 according to the CO-RADS and CT severity score systems. We compared findings of the CO-RADS AI system with readings of eight observers and with clinical assessments of the patients, including reverse transcription polymerase chain reaction (RT-PCR) test results.

Materials and Methods

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

Study Sample

We retrospectively included consecutive patients arriving at the emergency wards of an academic center and a large teaching hospital in the Netherlands in March and April of 2020 who 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 radiologic report were excluded. Additionally, patients from the teaching hospital were excluded if they were known to have COVID-19 (proved with RT-PCR testing) 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 (online).

Because RT-PCR testing may initially yield false-negative results, we considered patients to have COVID-19 if they had a positive RT-PCR test result or if their clinical presentation made COVID-19 the probable diagnosis. Criteria were the lack of an alternative diagnosis explaining the symptoms and admission to the intensive care unit 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 were used for model development, including 520 CT scans in total. CO-RADS scores were extracted from the radiologic reports. Scans with a CO-RADS score of 6, which signifies a positive RT-PCR result prior to imaging, were rescored independently by a chest radiologist with more than 30 years of experience (E.T.S.) who assigned CO-RADS scores of 1–5 to simulate an unknown RT-PCR status (n = 52). This observer was blinded to the original radiologic report and to all nonimaging data except 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 were set aside to verify the performance of the AI model. For each patient, at least one RT-PCR result was available within 5 days after CT scan acquisition. The earliest available scan of each patient was scored independently after CO-RADS classification by seven chest radiologists and one radiology resident (B.G., J.K., L.B., M.P., H.A.G., J.L.S., C. Schaefer-Prokop, T.v.R.V.) using a dedicated browser-based workstation (CIRRUS, Diagnostic Image Analysis Group, Nijmegen, the Netherlands) (available at *https://grand-challenge.org/reader-studies/*). The observers were familiar with the CO-RADS and CT severity scoring systems from interpreting at least 30 scans. Four of them (B.G., J.K., J.L.S., and T.v.R.V.) had less than 5 years of experience in reading

chest CT scans, and the others had up to 27 years. Any available later scans from these 105 patients were not used in this study. All observers were blinded to the RT-PCR test results and therefore could not assign a CO-RADS score of 6. Instead, they assigned CO-RADS scores of 1–5 based on their suspicion of COVID-19 pulmonary involvement. In addition, they semiquantitatively described the extent of parenchymal involvement per lobe using a predefined CT severity score on a six-point scale (0 = 0%, 1 = 1%–5%, 2 = 6%–25%, 3 = 26%–50%, 4 = 51%–75%, and 5 = >75%) (5).

External test set.—The data of all patients included at the teaching hospital were 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 score (the sum of the scores per lobe, as described previously) were extracted from the radiologic 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, version 1.1.1; 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 morphologic operations. Vessels and airways were removed using automatic methods. Lesion candidates in lobes not affected by COVID-19 following the radiologic report were removed. The remaining lesion candidates were reviewed by a certified image analyst with at least 1 year of experience in correcting segmentations of pulmonary structures on chest CT scans. The analyst corrected the delineations and added and removed lesions, as needed.

Automated CT Scoring

CT scans were scored automatically using three successively applied deep-learning algorithms. These performed pulmonary lobe segmentation and labeling, lesion segmentation and CT severity score prediction, and 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 pretrained on 4000 chest CT scans from the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease study (21) and was fine-tuned with 400 scans from the present study.

For CT severity score prediction, we trained a three-dimensional 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 three-dimensional inflated Inception architecture (23,24), which is a three-dimensional extension of the state-of-the-art Inception image classification architecture (25). The model was pretrained on ImageNet (26) and Kinetics (27) data sets and was trained with 368 CT scans from the 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 (online). The algorithm is freely accessible online (*https://grand-challenge.org/algorithms/corads-ai/*).

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 and the area under the receiver operating characteristic curve (AUC). Youden index was used to determine the optimal threshold. Nonparametric bootstrapping with 1000 iterations was used to calculate 95% CIs. To quantify agreement, linearly weighted κ values and classification accuracy were determined by comparing the predicted CO-RADS and CT severity scores to the median of all combinations of scores from seven observers. The agreement of the AI system in terms of the linearly weighted κ value 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 χ^2 tests (sex) and tested for differences in the sensitivity of the observers and algorithm at the specificity of the observers using McNemar tests. The significance level was .05. Analyses were performed with statistical software (R, version 3.6.2; R Foundation for Statistical Computing, Vienna, Austria) and Python, version 3.7.6 (scipy 1.5.0, sklearn 0.23.1, evalutils 0.2.3; Python Software Foundation, Wilmington, Del).

Results

Patient Characteristics

A total of 581 and 262 consecutive patients were included at the academic center and the teaching hospital, respectively. The training and development set comprised 520 scans of 476 patients from the academic center, and the internal test set comprised 105 scans of the remaining 105 patients. The external test set comprised 262 scans of 262 patients from
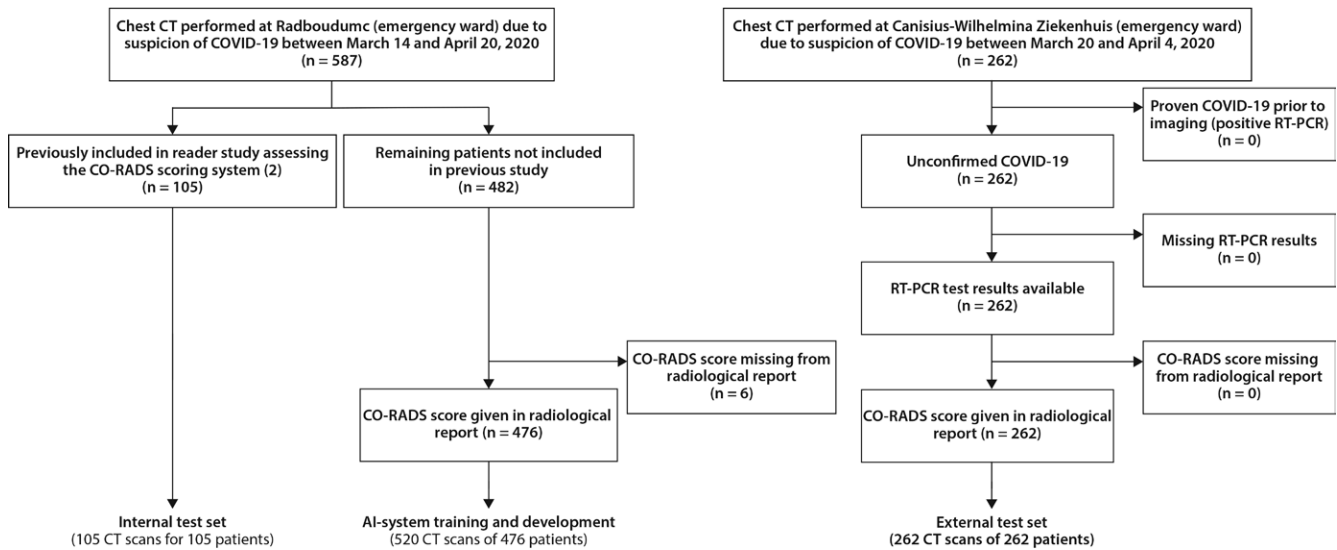
Figure 1: Flowchart shows patient inclusion in the training and test sets. Note that *n* refers to the number of patients. The number of CT images is higher in the training set, as several patients underwent multiple chest CT examinations during the inclusion period. However, in the test sets, only the earliest available scan for each patient is used. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019, RT-PCR = reverse transcription polymerase chain reaction.

Table 1: Characteristics of Training and Test Cohorts

Parameter	Training Set (N = 476)	Internal Test Set (N = 105)	External Test Set (N = 262)
Age	60 ± 16	62 ± 16*	64 ± 16[†]
Sex[‡]			
Male	267 (56)	61 (58)	154 (54)
Female	209 (44)	44 (42)	120 (46)
No. of CT scans per patient			
1	438 (92)	105 (100)	262 (100)
2	34 (7)	…	…
3	2 (0)	…	…
4	2 (0)	…	…

Note.—Data are numbers with percentages in parentheses or means ± standard deviations. *P* values correspond to comparisons of the distribution of characteristics in the training set and the respective test set.

* *P* = .20.

[†] *P* < .01.

[‡] *P* = .79 for the internal test set, *P* = .68 for the external test set.

Table 2: CO-RADS and CT Severity Scores according to Radiologic Reports or Radiologist

Parameter	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)

Note.—The radiologist reviewed scans with an initial CO-RADS score of 6 (proven COVID-19). CO-RADS data are numbers of CT scans with percentages in parentheses. CT-severity-score data are medians and interquartile ranges and were reported only for cases with CO-RADS scores greater than or equal to 3. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019.

the teaching hospital. Six patients were excluded because CO-RADS scores were missing from their radiologic reports (Fig 1). Patient characteristics for training and test sets are given in Table 1.

There were 58 patients with a clinical diagnosis of CO-VID-19 among the 105 patients in the internal test set (55%), and there were 179 such patients among the 262 patients in the external test set (68%). Of these patients, 53 of 58 (91%) and 145 of 179 (81%) had a positive RT-PCR result, whereas the remaining patients diagnosed with COVID-19 had one or multiple negative RT-PCR test results but were diagnosed with COVID-19 on the basis of their symptoms. Table 2 summarizes the distribution of

CO-RADS and CT severity scores according to the radiologic reports. The algorithm was executed successfully for all scans in the test sets. For the 105 scans in the internal test set, the median runtime of the algorithm was 212 seconds (range, 146–709), whereas the median reading time for radiologists was 82 seconds (range, 58–134 seconds).

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 because of 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%
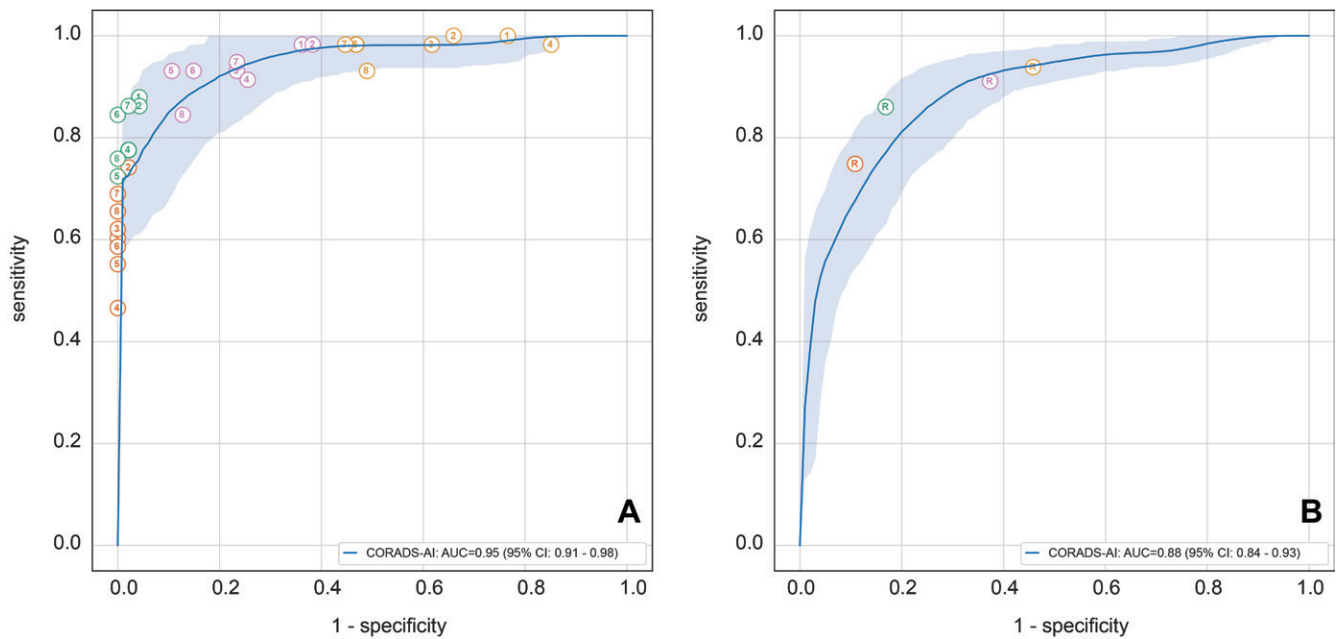
Figure 2: Receiver operating characteristic (ROC) curves for automatically predicted probability of a Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) score of 5 versus probability of a COVID-19 diagnosis. The ROC curve is based on the probability that the algorithm assigned a CO-RADS score of 5. The shaded area around the ROC curve reflects the 95% CI. *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 radiologic reports is shown for the external test set. Different colors indicate different cutoffs, in which patients were considered to be predicted as having COVID-19 if the observer assigned a CO-RADS score of 5 (orange), 4 or 5 (green), 3–5 (magenta), or 2–5 (yellow). COVID-19 diagnosis meant either a positive reverse transcription polymerase chain reaction (RT-PCR) test result or very high clinical suspicion of COVID-19, despite at least one negative RT-PCR test result. AUC = area under the ROC curve, CORADS-AI = CO-RADS artificial intelligence system.

± 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 Patients with COVID-19
In the internal test set, the algorithm distinguished between patients without and those with COVID-19 with an AUC of 0.95 (95% CI: 0.91, 0.98) on the basis of the probability of a CO-RADS score of 5 predicted by the algorithm. 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 receiver operating characteristic 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 sensitivity of the eight observers was 61.4% ± 7.9 (standard deviation) at a mean specificity of 99.7% ± 0.7, which was based on patients to whom they assigned a CO-RADS score of 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 three 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 score combinations from 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 8 × 105 = 840) of the patients and within one category of the seven observers' score in 80.5% (676 of 840). The remaining reader was in absolute agreement in 68.2% (573 of 840) of the patients and within one category of the other observers' score in 96.2% (808 of 840) of the patients. In the external test set, the AI algorithm score and the reference score were in absolute agreement in 64.12% (168 of 262) of the patients and within one category of one another in 85.50% (224 of 262) of the patients. The cross-tabulated results are given in Tables E1 and E2 (online).

There was moderate to substantial agreement between AI-predicted CO-RADS scores and the observers' score according to the linearly weighted κ value (Table 4). For the internal test set, the mean (± standard deviation) κ value was 0.60 ± 0.01 for the AI system and 0.79 ± 0.04 for the left-out observer ($P < .001$ for all observers). For the external test set, the κ value was 0.69 (95% CI: 0.63, 0.75) for the AI system.

CT Severity Score Prediction
Because 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

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

Parameter	Specificity	Observer Sensitivity	AI Sensitivity	P Value
Internal test set (n = 105)				
CO-RADS 5				
Observer 1	100	60.3	71.9 (58.7, 84.8)	.143
Observer 2	97.9	74.1	73.0 (59.3, 88.3)	>.999
Observer 3	100	62.1	71.9 (58.7, 84.8)	.238
Observer 4	100	46.6	71.9 (58.7, 84.8)	<.001
Observer 5	100	55.2	71.9 (58.7, 84.8)	.064
Observer 6	100	58.6	71.9 (58.7, 84.8)	.143
Observer 7	100	69	71.9 (58.7, 84.8)	>.999
Observer 8	100	65.5	71.9 (58.7, 84.8)	.607
CO-RADS 4 + 5				
Observer 1	95.7	87.9	76.1 (61.2, 92.3)	.146
Observer 2	95.7	86.2	76.1 (61.2, 92.3)	.302
Observer 3	97.9	77.6	73.0 (59.3, 88.3)	.791
Observer 4	97.9	77.6	73.0 (59.3, 88.3)	.791
Observer 5	100	72.4	71.9 (58.7, 84.8)	>.999
Observer 6	100	84.5	71.9 (58.7, 84.8)	.118
Observer 7	97.9	86.2	73.0 (59.3, 88.3)	.180
Observer 8	100	75.9	71.9 (58.7, 84.8)	.815
CO-RADS 3 + 4 + 5				
Observer 1	63.8	98.3	97.2 (90.9, 100.0)	>.999
Observer 2	61.7	98.3	97.5 (91.4, 100.0)	>.999
Observer 3	76.6	93.1	93.6 (83.1, 100.0)	>.999
Observer 4	74.5	91.4	94.5 (84.7, 100.0)	.549
Observer 5	89.4	93.1	85.6 (69.2, 96.6)	.607
Observer 6	85.1	93.1	89.0 (75.8, 98.2)	.791
Observer 7	76.6	94.8	93.6 (83.1, 100.0)	>.999
Observer 8	87.2	87.3	87.3 (72.0, 96.9)	.815
CO-RADS ≥ 2				
Observer 1	23.4	100	99.1 (94.7, 100.0)	>.999
Observer 2	34	100	98.2 (93.8, 100.0)	>.999
Observer 3	38.3	98.3	98.2 (93.8, 100.0)	>.999
Observer 4	14.9	98.3	99.9 (98.1, 100.0)	.754
Observer 5	53.2	98.3	98.1 (93.1, 100.0)	>.999
Observer 6	53.2	98.3	98.1 (93.1, 100.0)	>.999
Observer 7	55.3	98.3	98.0 (92.9, 100.0)	>.999
Observer 8	51.1	93.1	98.1 (93.1, 100.0)	.503
External test set (n = 262)				
CO-RADS 5	89.2	74.9	67.6 (54.6, 81.7)	.035
CO-RADS 4 + 5	83.1	86	77.1 (63.1, 88.6)	.020
CO-RADS 3 + 4 + 5	62.7	91.1	92.5 (86.3, 96.7)	.585
CO-RADS ≥ 2	54.2	93.9	94.2 (90.1, 97.7)	>.999

Note.—Data are percentages or percentages, with 95% CIs in parentheses. P values correspond to comparisons of the sensitivity of the observer with that of the artificial intelligence (AI) algorithm. CO-RADS = COVID-19 Reporting and Data System, COVID-19 = coronavirus disease 2019

internal test set. The predicted score was in absolute agreement with the median score of all score combinations from seven of the eight observers in 17.2% (80 of 8 × 58 positive patients = 464) and deviated by not more than one point per lobe (ie, five points in total) in 94.0% (436 of 464 patients). In the external test set, the radiologic reports contained severity scores for 163 of the 179 patients with COVID-19. 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' score according to the linearly weighted κ value (Table 4). For the internal test set, the mean κ value was 0.54 ± 0.01 for the AI system and 0.77 ± 0.03 for the left-out observer (P < .001 for all observers). For the external test set, the κ value was 0.49 (95% CI: 0.41, 0.56) for the AI system.

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

Parameter	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)	<.001
Observer 2	0.78 (0.71, 0.86)	0.59 (0.49, 0.70)	<.001
Observer 3	0.85 (0.80, 0.90)	0.60 (0.50, 0.69)	<.001
Observer 4	0.72 (0.65, 0.79)	0.60 (0.50, 0.71)	.018
Observer 5	0.74 (0.66, 0.81)	0.61 (0.51, 0.71)	.032
Observer 6	0.84 (0.79, 0.90)	0.61 (0.52, 0.71)	<.001
Observer 7	0.82 (0.77, 0.88)	0.58 (0.48, 0.68)	<.001
Observer 8	0.79 (0.73, 0.86)	0.61 (0.51, 0.70)	.002
Total CT severity score			
Observer 1	0.80 (0.74, 0.85)	0.53 (0.44, 0.63)	<.001
Observer 2	0.74 (0.68, 0.81)	0.53 (0.43, 0.62)	<.001
Observer 3	0.78 (0.71, 0.84)	0.54 (0.45, 0.63)	<.001
Observer 4	0.76 (0.69, 0.84)	0.53 (0.44, 0.63)	<.001
Observer 5	0.80 (0.74, 0.87)	0.54 (0.45, 0.63)	<.001
Observer 6	0.71 (0.64, 0.79)	0.55 (0.45, 0.64)	.004
Observer 7	0.81 (0.75, 0.87)	0.53 (0.44, 0.63)	<.001
Observer 8	0.73 (0.65, 0.80)	0.55 (0.46, 0.64)	.002

Note.—Data are linearly weighted κ scores with 95% CIs in parentheses. *P* values correspond to comparisons of the κ scores of the left-out observer and AI system using the median of the other seven observers as the reference. AI = artificial intelligence, CO-RADS = Coronavirus Disease 2019 Reporting and Data System.
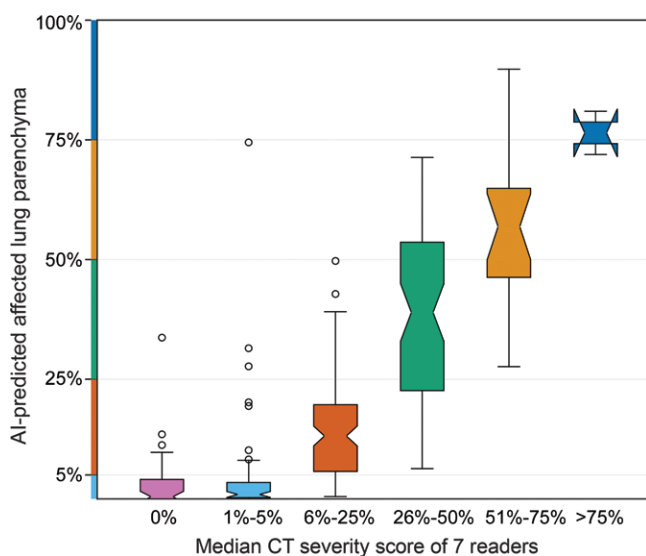
Figure 3: CT severity score predictions versus median of observer scores. The distribution of the percentage of affected lung parenchyma per lobe according to the automatic lesion (affected volume) and lobe segmentations (total volume) for the internal test set are shown as box plots. The notch in each box plot illustrates the 95% CI around the median. The CT severity score cutoffs are marked on the y-axis. AI = artificial intelligence.

Representative examples of lobe segmentation results, CO-RADS score predictions, and CT severity score predictions with corresponding pulmonary lesions are shown in Figures 4–6 and Appendix E3 (online).

Discussion

Artificial intelligence (AI) might be helpful in interpreting CT scans of patients highly suspected of having coronavirus disease 2019 (COVID-19), especially when it produces standardized output with which radiologists and other health care providers are familiar. In this study, we evaluated the performance of an AI system for automated scoring of chest CT scans of patients suspected of having COVID-19 on the basis of the COVID-19 Reporting and Data System (CO-RADS) and CT severity score classifications. This system identified patients with COVID-19 with high diagnostic performance, achieving an area under the curve (AUC) of 0.95 in the test set with sensitivity and specificity similar to those of the eight observers and achieving an AUC of 0.88 in the external test set. The automated CO-RADS scoring was in good agreement with that of the observers, although agreement was significantly higher between observers (mean κ value, 0.60 for AI agreement vs 0.79 for interobserver agreement; $P < .001$). Likewise, the automated CT severity score agreed well with observer scoring, with a mean κ value of 0.54 ± 0.01, which was also lower than the agreement among the observers (κ = 0.77 ± 0.03, $P < .001$). 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 in which automatic measurements were more than 10 points higher than the reference, underlying causes were severe motion artifacts in three cases and opacifications caused by aspiration pneumonia in one case. This emphasizes 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 CO-VID-19. Shi et al (29) and Ito et al (30) provide overviews of the proposed approaches. Most studies analyze small data
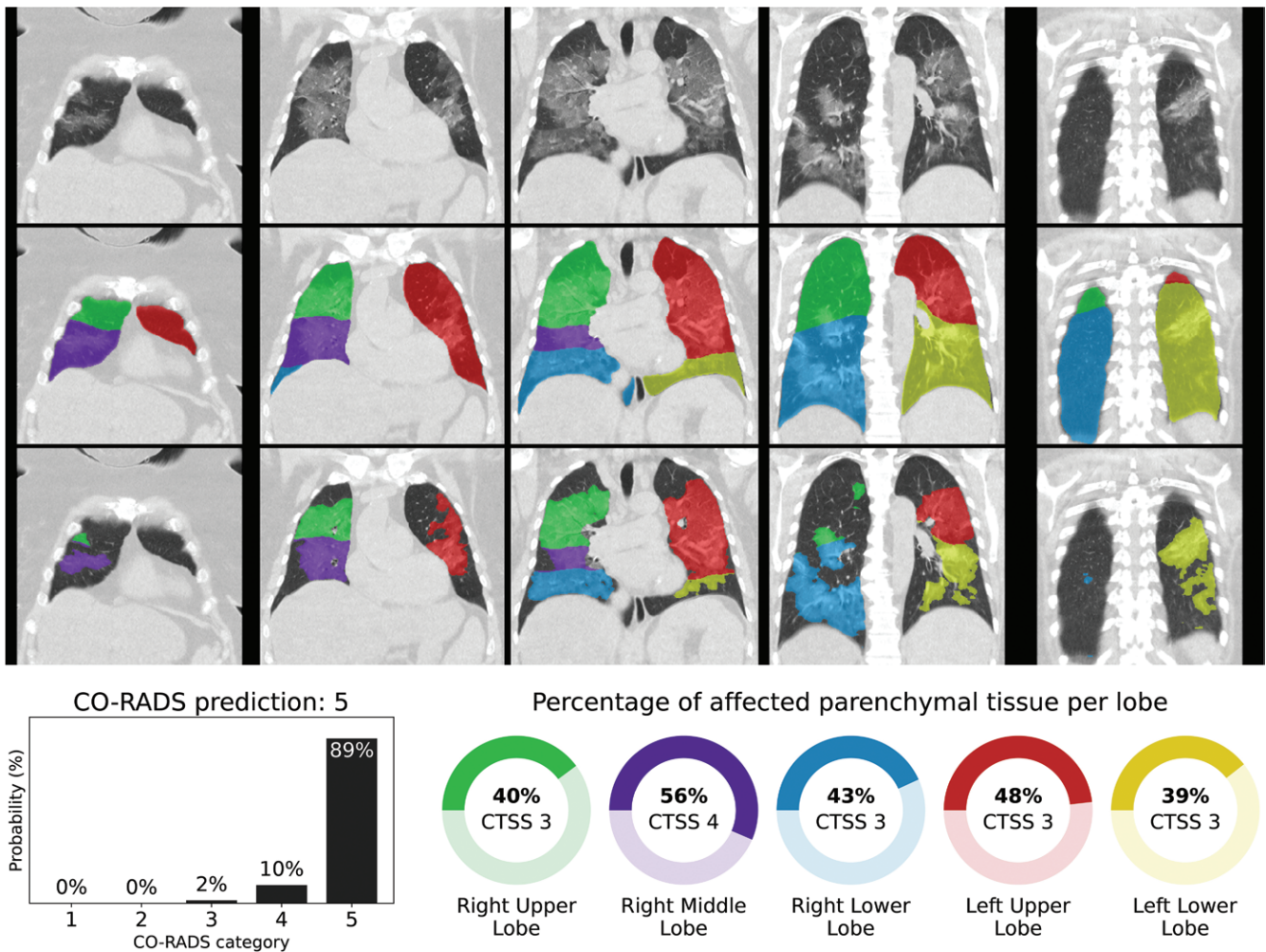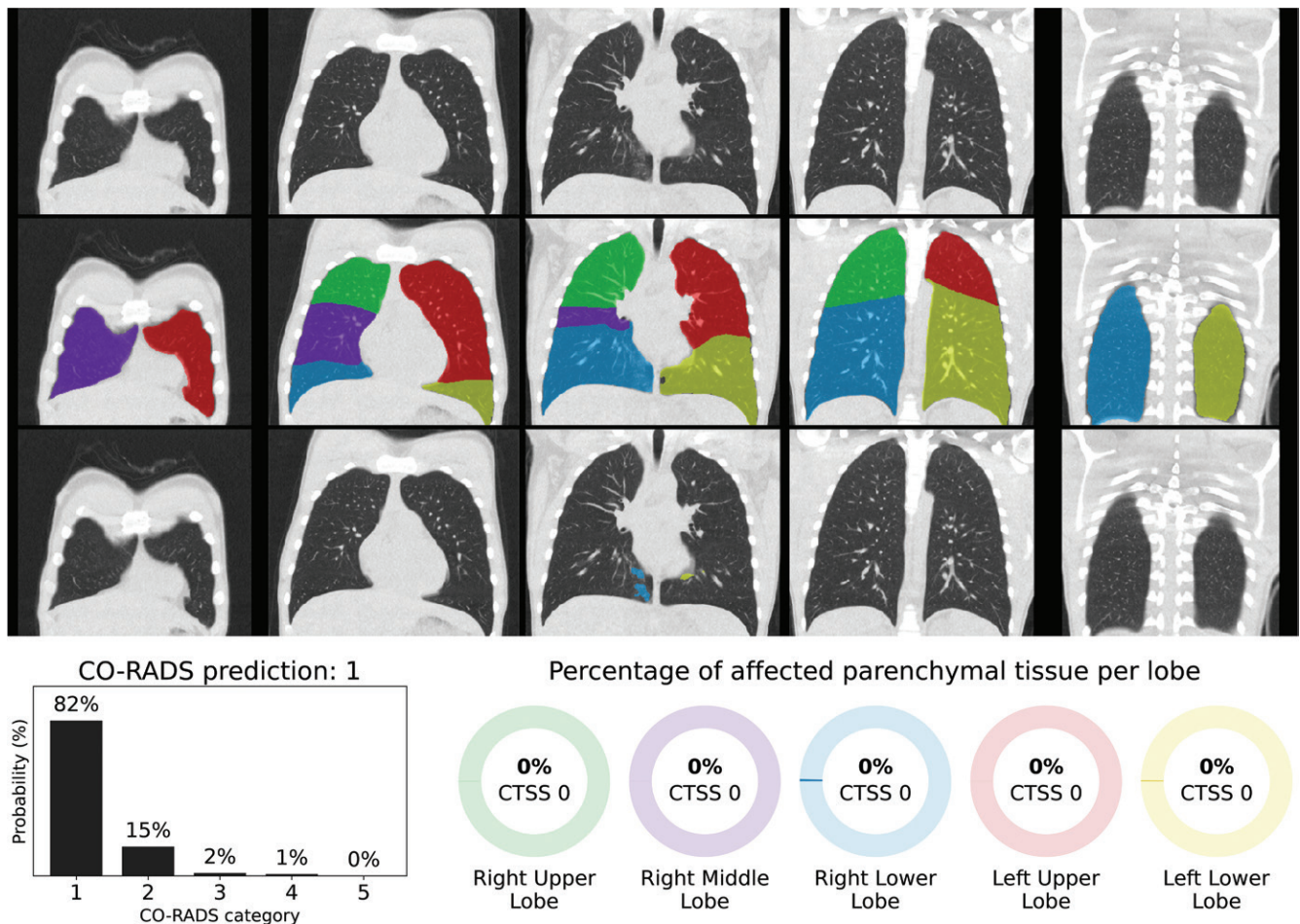
Figure 4: Coronavirus Disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–positive case with extensive parenchymal involvement. Scans from a 73-year-old woman with a positive reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left) as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 3 (three observers), 4 (one observer), and 5 (four observers).

sets and employ two-dimensional neural network architectures on axial sections. One of the first large studies by Li et al (9) with such a two-dimensional 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 took a two-dimensional approach and showed in 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 to be 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 decreased 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 two-dimensional approach (AUC = 0.95), reducing to an AUC of 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 using CT scans and showed that this may lead to better performance on external data.

Mei et al (18) developed a COVID-19 diagnostic AI system and tested it on a held-out set of 279 patients. They followed a two-dimensional approach similar to 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 morphologic features with other non-COVID-19–related diseases, our study could be advanced with an AI analysis of imaging, clinical signs, and
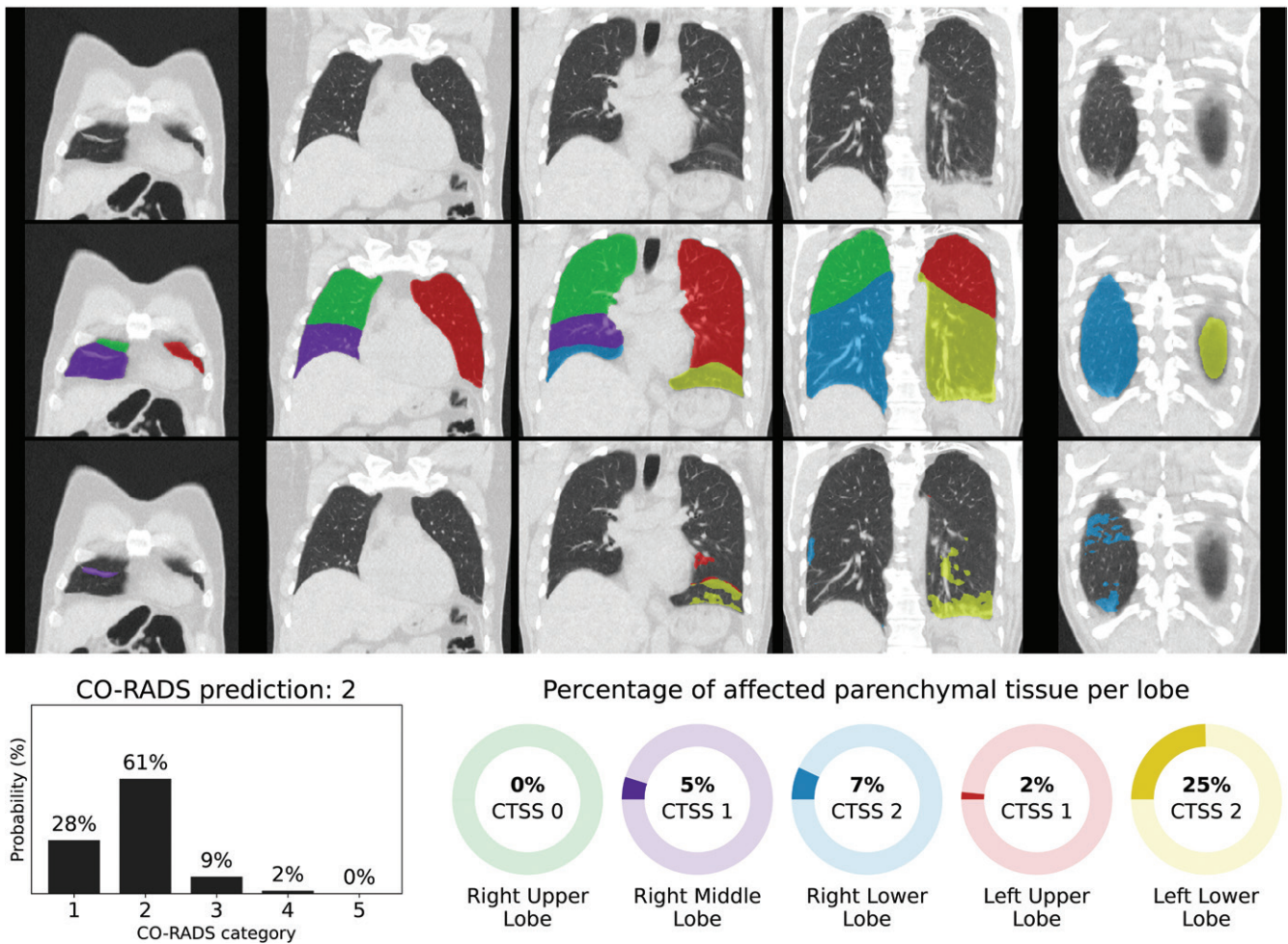
Figure 5: Coronavirus disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–positive case with little parenchymal involvement. Scans from an 18-year-old man with a positive reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 1 (two observers), 2 (five observers), and 3 (one observer).

routine laboratory parameters. There are other studies that have 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 these systems therefore provide an unstandardized uncalibrated output. The lack of explicability 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 in which the accuracy of automated lobar segmentation and lesion segmentation on CT scans of patients suspected of having COVID-19 was quantitatively evaluated and compared with human readers.

Our study had several limitations. First, we trained the AI system with data from one medical center. More training is needed with multicenter data sets. Second, the study sample represented a population of patients who arrived in a high-prevalence situation at our hospital and underwent

chest CT examinations because they were suspected of having COVID-19. There was only a limited number of patients with extensive preexisting lung disease in the training set. Third, inclusion of the study group took place after the influenza and respiratory syncytial 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 data set 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 patients without COVID-19.

In conclusion, our study demonstrates that an artificial intelligence (AI) system can identify patients with coronavirus disease 2019 (COVID-19) on the basis of unenhanced chest CT images with diagnostic performance comparable to that of radiologic observers. It is noteworthy that the algorithm was trained to adhere to the COVID-19 Reporting and Data System categories and is thus directly interpretable

Figure 6: Coronavirus disease 2019 (COVID-19) Reporting and Data System (CO-RADS) and CT severity score (CTSS) predictions for a COVID-19–negative case. Scans from a 54-year-old man with a negative reverse transcription polymerase chain reaction test result are shown. Nonenhanced CT scans in the coronal view (top row) overlaid with the automatic lobe segmentation (middle row) and the detected areas of abnormal parenchymal lung tissue (bottom row) are shown. This figure also shows the probabilities that the artificial intelligence model assigned to each CO-RADS category (bottom left), as well as the computed percentages of affected parenchymal tissue and the corresponding CT severity score per lobe (bottom right). The eight observers assigned this case CO-RADS scores of 1 (three observers), 2 (three observers), and 3 (two observers).

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 also be used to quantify lung damage during patient follow-up.

References

1. Yang W, Sirajuddin A, Zhang X, et al. The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). Eur Radiol 2020;30(9):4874–4882.
2. Prokop M, van Everdingen W, van Rees Vellinga T, et al. CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19—definition and evaluation. Radiology 2020;296(2):E97–E104.
3. Simpson S, Kay FU, Abbara 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—secondary publication. J Thorac Imaging 2020;35(4):219–227.
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. Eur Radiol 2020;30(9):4930–4942.
5. Li K, Wu J, Wu F, et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Invest Radiol 2020;55(6):327–331.
6. Chang YC, Yu CJ, Chang SC, et al. Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT. Radiology 2005;236(3):1067–1075.
7. Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia. Appl Intell doi: 10.1007/s10489-020-01714-3. Published online April 22, 2020. Accessed May 22, 2020.
8. Song Y, Zheng S, Li L, et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. MedRxiv 10.1101/2020.02.23.20026930v1 [preprint]. Posted February 25, 2020. Accessed May 22, 2020.
9. Li L, Qin L, Zeguo X, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 2020;296(2):E65–71. Published online March 19, 2020. Accessed May 22, 2020.
10. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). MedRxiv 10.1101/2020.02.14.20023028v5. Posted April 24, 2020. Accessed May 27, 2020.
11. Jin S, Wang B, Xu H, et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system in four weeks. MedRxiv 10.1101/2020.03.19.20039354v1 [preprint]. Posted March 23, 2020. Accessed May 27, 2020.
12. Bai HX, Wang R, Xiong Z, et al. AI augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT. Radiology 2020;296(3):E156–E165.
13. Ouyang X, Huo J, Xia L, et al. Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia. IEEE Trans Med Imaging 2020;39(8):2595–2605.
14. Wang J, Bao Y, Wen Y, et al. Prior-attention residual learning for more discriminative COVID-19 screening in CT images. IEEE Trans Med Imaging 2020;39(8):2572–2583.
15. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. ArXiv 2003.05037 [preprint]. Posted March 10, 2020. Accessed May 22, 2020.
16. Shan F, Gao Y, Wang J, et al. Lung infection quantification of COVID-19 in CT images with deep learning. ArXiv 2003.04655 [preprint] https://arxiv.org/abs/2003.04655. Posted March 10, 2020. Accessed May 22, 2020.
17. Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020;181(6):1423–1433.e11.
18. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26(8):1224–1228.
19. Kundu S, Elhalawani H, Gichoya JW, Kahn CE. How might AI and chest imaging help unravel COVID-19's mysteries? Radiol Artif Intell 2020;2(3):e200053.
20. Xie W, Jacobs C, Charbonnier JP, van Ginneken B. Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans. IEEE Trans Med Imaging 2020;39(8):2664–2675.
21. Regan EA, Hokanson JE, Murphy JR, et al. Genetic Epidemiology of COPD (COPDGene) study design. COPD 2010;7(1):32–43.
22. Isensee F, Jäger PF, Kohl SAA, Petersen J, Maier-Hein KH. Automated design of deep learning methods for biomedical image segmentation. ArXiv 1904.08128 [preprint]. Posted April 17, 2019. Accessed May 22, 2020.
23. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017; 4724–4733.
24. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–961. [Published correction appears in Nat Med 2019; 25(8):1319.]
25. Szegedy C, Wei L, Yangqing J, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2015; 1–9.
26. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–252.
27. Li A, Thotakuri M, Ross DA, Carreira J, Vostrikov A, Zisserman A. The AVA-kinetics localized human actions video dataset. ArXiv 2005.00214 [preprint]. Posted May 1, 2020. Accessed May 1, 2020.
28. Gietema HA, Müller NL, Fauerbach PV, et al. 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.
29. Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev Biomed Eng doi: 10.1109/RBME.2020.2987975. Published online April 16, 2020. Accessed July 4, 2020.
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;26(5):443–448.
31. Xu Y, Ma L, Yang F, et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. MedRxiv 10.1101/2020.05.10.20096073v2 [preprint]. Posted May 19, 2020. Accessed July 4, 2020.
32. Dofferhoff ASM, Swinkels A, Sprong T, et al. Diagnostic algorithm for COVID-19 at the ER [in Dutch]. Ned Tijdschr Geneeskd 2020;164:D5042.
33. Kurstjens S, van der Horst A, Herpers R, et al. Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing. Clin Chem Lab Med 2020;58(9):1587–1593.