Radiology: Cardiothoracic Imaging

Comparison of Chest CT Grading Systems in Coronavirus Disease 2019 (COVID-19) Pneumonia

Shohei Inui, MD[1, 2]; Ryo Kurokawa, MD, PhD[1]; Yudai Nakai, MD, PhD[1]; Yusuke Watanabe, MD, PhD[1]; Mariko Kurokawa, MD[3]; Keita Sakurai, MD, PhD[4]; Akira Fujikawa, MD, PhD[2]; Hiroaki Sugiura, MD, PhD[5]; Takuya Kawahara, PhD[6]; Soon Ho Yoon, MD, PhD[7]; Yasuhide Uwabe, MD, PhD[8]; Yuto Uchida, MD, PhD[9]; Wataru Gonoi, MD, PhD[1, *]; Osamu Abe, MD, PhD[1]

1. Department of Radiology, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8655, Japan
2. Department of Radiology, Japan Self-Defense Forces Central Hospital, 1-2-24, Ikejiri, Setagaya-ku, Tokyo, 154-0001, Japan
3. Department of Radiology, Tokyo Metropolitan Cancer and Infectious Diseases Center Komagome Hospital, 3-18-22, Honkomagome, Bunkyo-ku, Tokyo, 113-8677, Japan
4. Department of Radiology, National Center for Geriatrics and Gerontology, 7-430, Morioka-cho, Obu, Aichi, 474-8511, Japan
5. Department of Radiology, National Defense Medical College, 3-2, Namiki, Tokorozawa-shi, Saitama, 359-8513, Japan
6. Clinical Research Promotion Center, The University of Tokyo Hospital, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8655, Japan
7. Department of Radiology, Seoul National University College of Medicine, Seoul National University Hospital, 101 Daehak-ro, Chongno-gu, Seoul 03080, Republic of Korea
8. Department of Respiratory Medicine, Japan Self-Defense Forces Central Hospital, 1-2-24, Ikejiri, Setagaya-ku, Tokyo, 154-0001, Japan
9. Department of Neurology and Neuroscience, Nagoya City University Graduate School of

Medical Sciences, 1, Kawasumi, Mizuho-ku, Nagoya, 467-8601, Japan

*Corresponding author

Correspondence to:

Wataru Gonoi, MD, PhD

Department of Radiology, Graduate School of Medicine, The University of Tokyo

7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8655, Japan

Email: watapi-tky@umin.net

Funding:

No funding is provided in this study.

Disclosure of Conflicts of Interest:

No financial conflicts of interest to disclose with regard to this study.

Key Points

・The level of suspicion of COVID-19 correlated with the RT-PCR positive rate except for the "negative for pneumonia" classifications (Spearman's coefficient: ρ=1.0, *P*=<.001 for all the systems).

・Average AUCs were as follows: CO-RADS, 0.84 (95% confidence interval: 0.83–0.85): COVID-RADS, 0.80 (0.78-0.81): the RSNA expert consensus statement, 0.81 (0.79-0.82): and the BSTI guidance statement, 0.84 (0.82-0.86).

・Average Cohen's kappa in all observers was as follows: CO-RADS, 0.62 (95% confidence interval: 0.58–0.66): COVID-RADS, 0.63 (0.58-0.68): the RSNA expert consensus statement, 0.63 (0.57–0.69): and the BSTI guidance statement, 0.61 (0.58-0.64).

Summary statement:

The CT grading systems of COVID-19 provided a reasonable diagnostic performance with AUCs ranging 0.80-0.84 and good interobserver agreement with Cohen's kappa values ranging 0.61-0.63, along with positive correlations between RT-PCR positive rates and categories except for the "negative for pneumonia" classifications.

Abbreviations:

CoV: Coronavirus

RT-PCR: reverse transcription-polymerase chain reaction

CT: computed tomography

GGO: Ground-glass opacity

CO-RADS: the COVID-19 Reporting and Data System

COVID-RADS: the COVID-19 imaging reporting and data system

BSTI: British Society of Thoracic Imaging

Abstract

Purpose: To compare the performance and interobserver agreement of the COVID-19 Reporting and Data System (CO-RADS), the COVID-19 imaging reporting and data system (COVID-RADS), the RSNA expert consensus statement, and the British Society of Thoracic Imaging (BSTI) guidance statement.

Materials and Methods: In this case-control study, total of 100 symptomatic patients suspected of having COVID-19 were included: 50 patients with COVID-19 (59±17 years, 38 men) and 50 patients without COVID-19 (65±24 years, 30 men). Eight radiologists independently scored chest CT images of the cohort according to each reporting system. The area under the receiver operating characteristic curves (AUC) and interobserver agreements were calculated and statistically compared across the systems.

Results: A total of 800 observations were made for each system. The level of suspicion of COVID-19 correlated with the RT-PCR positive rate except for the "negative for pneumonia" classifications in all the systems (Spearman's coefficient: ρ=1.0, *P*=<.001 for all the systems). Average AUCs were as follows: CO-RADS, 0.84 (95% confidence interval, 0.83–0.85): COVID-RADS, 0.80 (0.78–0.81): the RSNA statement, 0.81 (0.79–0.82): and the BSTI statement, 0.84 (0.812-0.86). Average Cohen's kappa across observers was 0.62 (95% confidence interval, 0.58–0.66), 0.63 (0.58–0.68), 0.63 (0.57–0.69), and 0.61 (0.58-0.64) for CO-RADS, COVID-RADS, the RSNA statement and the BSTI statement, respectively. CO-RADS and the BSTI statement outperformed COVID-RADS and the RSNA statement in diagnostic performance (*P*=.<.05 for all the comparison).

Conclusions: CO-RADS, COVID-RADS, the RSNA statement and the BSTI statement provided reasonable performances and interobserver agreements in reporting CT findings of COVID-19.

Introduction

During the ongoing pandemic of coronavirus disease 2019 (COVID-19), the prompt diagnosis is crucial to achieve swift and optimal clinical decision making and judge the precaution level necessary on admission to help prevent nosocomial infection in the hospital. Various evidence has documented that early diagnosis and intervention are associated with a better prognosis [1]. The gold standard diagnostic method for COVID-19 is reverse transcription-polymerase chain reaction (RT-PCR) that directly quantifies viral load from a nasopharyngeal swab, sputum, or endotracheal lavage. However, the sensitivity of this method is unclear as false-negative results have been reported in patients with insufficient specimen or those in the initial stage of infection [2]. The turnaround time is also long ranging from a few hours to days. Although chest CT is currently not recommended for routine screening purposes, it provides valuable information serving as a supplementary diagnostic tool of COVID-19 pneumonia especially in circumstances in which RT-PCR tests are not sufficiently available or in patients in whom the possibility of false negative results is suspected, or clinical decisions are required before the PCR test results become available.

With the accumulation of recent publications clarifying the radiological appearance of COVID-19, various attempts have been made to standardize reporting of chest CT for suspected COVID-19. The British Society of Thoracic Imaging (BSTI) proposed Guidance for the Reporting Radiologist as a diagnostic framework of COVID-19 from chest CT and radiograph [3]. The recent RSNA expert consensus statement on reporting advocates a standard nomenclature and imaging classification for COVID-19 pneumonia made up of four categories (typical appearance, indeterminate appearance, atypical appearance, and negative for pneumonia) [4]. A working group of the Dutch Radiological Society devised the COVID-19 Reporting and Data System (CO-RADS) to facilitate the advances in and worldwide dissemination of COVID-19 related information and tools [5]. Another group of researchers devised a different structured reporting system based on a review of 37 published papers on the chest CT findings of COVID-19 entitled the COVID-19 imaging reporting and data system (COVID-RADS) that divides the CT findings into five categories [6].

The published CT grading systems of chest CT findings in COVID-19 patients may facilitate both making the radiological diagnosis and smooth communication among professionals in other fields, and their applicability and validity in the clinical practice was recently reported in several studies [7, 8, 9, 10]. However, no studies have yet directly compared the diagnostic performances and interobserver agreement between them. This prompted us to undertake the present study to validate the performance and interobserver agreement of four sets of CT grading systems, including the COVID-19 Reporting and Data System (CO-RADS), the COVID-19 imaging reporting and data system (COVID-RADS), the RSNA expert consensus statement and the BSTI guidance statement.

Materials and Methods

This study was conducted with the approval of our institutional ethics review board. Written informed consent was waived due to the retrospective nature of the study. The privacy of all patients was protected in full.

Sample Size Calculation and Study Population

Patient backgrounds were standardized by applying the following inclusion criteria: (1) presentation to the outpatient or emergency department of a single institution from January 30 to June 30, 2020, (2) suspected of COVID-19 because of the presence of symptoms suggestive of pneumonia (i.e. fever (>37.5℃) and at least one of the following symptoms; cough, dyspnea, tachypnea, or hypoxemia), (3) having undergone RT-PCR examination, (4) acquisition of chest CT within 5 days of the initial RT-PCR test. Patients were classified as COVID-19 or non-COVID-19 if they tested positive or negative respectively on RT-PCR at least one time. Those who tested negative on the initial RT-PCR but were on a high clinical suspicion of COVID-19 underwent repeat RT-PCR and categorized as COVID-19 positive if repeat RT-PCR tested positive. Those who tested negative on the initial RT-PCR and were not having a high clinical suspicion of COVID-19 or proved to carry other etiologies did not necessarily undergo repeat RT-PCR.

Cases included in a previous publication were excluded from the current study based on the following grounds: (1) the previous publication was used in the process of developing two of the sets of criteria (the RSNA expert consensus statement and COVID-RADS), (2) those included in the previous publication were cases from mass infection cohort under special circumstances, and (3) the purpose of this study was to compare the CT grading systems in usual clinical settings that mostly comprises community-acquired infection with COVID-19. Furthermore, patients with COVID-19 were randomly selected and excluded to adjust the sample size. Fifty patients with COVID-19 and 50 patients without COVID-19 were finally included as case and control subjects, respectively. Flow chart illustrating the patient population was summarized in Figure 1.

Clinical Data

Medical records were reviewed for the clinical and imaging findings of patients. The following data were extracted from the medical records: demographic data, presence or absence of smoking history, underlying comorbidities, symptoms and signs, and duration from onset to CT.

Chest CT Acquisition

Non-enhanced chest CT was performed using a 6-row multi-detector CT unit (SOMATOM Emotion 6 scanner; Siemens, Tokyo, Japan) on admission with the following parameters: tube voltage, 130 kVp; effective current 95 mA; collimation, 6×2 mm, helical pitch, 1.4, field of view, 38 cm; matrix size, 512×512. A 1.0-mm gapless section was reconstructed before being reviewed on the picture archiving and communication system monitor.

CT Image Interpretation

The comparison of four sets of CT grading systems were summarized in Table 1. Eight general radiologists (W.G., Y.W., Y.N., R.K., M.K., K.S., H.S., A.F.) served as observers from 5 hospitals in Japan. Four observers had more than 10 years of experience and four less than 10 years. Because none had experience with CO-RADS, COVID-RADS, the RSNA expert consensus statement or the

BSTI guidance statement prior to this study, we held a practice review session and consensus meeting to review the statements before the initiation of the experiments. In the practice session, each reader independently scored 40 sample cases of COVID-19, which were not included in this study and recorded the reasons for grading and uncertainty (if any). In the following consensus meeting, we decided on a single consensus grading for each case by majority voting followed by a review of cases of interobserver disagreements and listed unclear/unspecified components in each item of the grading systems (Table 2). Based on these results, we added some supplementary remarks regarding interpretation of criteria and sample CT patterns to facilitate CT grading without confusion or misclassification as follows; (1) added conjunction (and/or) for each item of the criterion, (2) confirmed interpretation of the descriptions regarding laterality (i.e. each category was interpreted as "either unilateral or bilateral" if otherwise specified), (3) confirmed categorization of frequently encountered differential diagnosis of COVID-19 for each criterion (e.g. interstitial pulmonary edema falls into COVID-RADS 1 with only interstitial septal thickening and/or pleural effusion; COVID-RADS 2A when accompanied by peribronchial edema; and COVID-RADS 2B when progressed to pulmonary alveolar edema accompanying GGO and pleural effusion), (4) created sample CT patterns of CO-RADS with regards to its categorization of GGO.

Randomization was stratified by the patient's disease status (patients with COVID-19 vs. patients without COVID-19). All patient information was removed from the data and observers were blinded to all clinical information, including RT-PCR results. Each observer independently scored the four criteria using an original document of each criterion with the above-described additional notes and recorded all data using a spreadsheet prepared in advance (Microsoft, Redmond, WA, USA). Grading was conducted in four separate sessions designated for each criterion, in which the observers provided a single grading for one case (i.e. in the CO-RADS session, the observers provided only CO-RADS gradings, and in the order of CO-RADS session, COVID-RADS session, RSNA session, and BSTI session). The order of the cases was shuffled between sessions.

Statistical Analysis

The AUC was calculated for each sets of criteria for each of the observers. Using the RT-PCR results as the gold standard of COVID-19 diagnosis, the AUC was used to assess the performance of each of the three sets of criteria. Mean AUC across observers and 95% confidence interval (CI) were calculated. Last, for each criterion, the average percentage of patients assigned to each category, including 95% CI, was determined. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated by setting different cut-off points for each criterion.

Interobserver agreement was quantified using three types of kappa values (i.e. Fleiss' kappa, Cohen's kappa, and Light's kappa [10]) calculated across observers. In comparison to the original article, Cohen's kappa values were obtained by comparing the scores of each observer to the median of the remaining seven observers [5]. Overall agreement was quantified using Fleiss' kappa and Light's kappa. The degree of interobserver agreement was considered with the following interval of kappa: 0–0.20 for poor, 0.21–0.40 for fair, 0.41–0.60 for moderate, 0.61–0.80 for good, and 0.81–1.00 for excellent agreement.

Statistical analysis was done using the JMP statistical software program (JMP Pro, version 15.0.0; SAS, Cary, NC, USA), R software (R version 3.6.2, The R Foundation for Statistical Computing, Vienna, Austria). Quantitative variables were expressed as mean ± standard deviation (range) or median and interquartile range based on the normality of data. Categorical variables were presented as the percentage of the total. The comparisons of quantitative variables were evaluated using a non-paired t-test or Mann-Whitney U-test and categorical data using the Pearson $\chi 2$ test. The comparisons of AUC and kappa-values were conducted using one-way repeated measures analysis of variance, according to the normal distribution assessed by the Shapiro-Wilk test, and post hoc family-wise error correction for multiple comparisons with paired-t-test. All *P* values correspond to two-sided tests and the statistical significance level was set at Holm-Bonferroni-corrected $P < .05$.

Results

Clinical Findings

Demographics and clinical characteristics of the study population are summarized in Table 3. The study population comprised 100 patients, 50 patients with COVID-19 (38 men; mean age, 59 years ±17; range, 18–86) and 50 patients without COVID-19 (30 men; mean age, 65 years ±24, range; 17-100). There was no statistical significance between these groups in age, sex, disease duration, smoking history, or presence of comorbidities.

Diagnostic performances of the criteria

Eight observers scored 100 patients, making for a total of 800 observations for each criterion. The probability of COVID-19 diagnosis of each category of each set of criteria was summarized in Table 4. The level of suspicion on COVID-19 correlated with the RT-PCR positive rate except for being negative for pneumonia in all the systems (Spearman's coefficient: $\rho = 1.0$, $P < .001$ for all the systems). The diagnostic performance and interobserver agreements of each set of the criterion were summarized in Table 5. Average AUCs with 95% CI for each criterion were as follows; CO-RADS, 0.84 (0.83–0.85): COVID-RADS, 0.80 (0.78–0.81): RSNA grading system, 0.81 (0.79–0.82): and BSTI grading system, 0.84 (0.82-0.86). The AUC values were significantly higher in CO-RADS and BSTI grading system vs. COVID-RADS (vs CO-RADS, P =.0087 and vs BSTI grading system, P =.0033) and RSNA grading system (vs CO-RADS, P =.0097 and vs BSTI grading system, P =.0019). AUC values were not statistically significant in either of the comparisons between CO-RADS and BSTI grading system or COVID-RADS and the RSNA grading system.

The sensitivity, specificity, PPV, and NPV of each set of criteria was summarized in Table 6. For CO-RADS, the sensitivity, specificity, PPV and NPV was as follows: CO-RADS 5, sensitivity 64.5% (258/400), specificity 89.0% (356/400), PPV 85.4% (258/302), and NPV 71.5% (356/498); CO-RADS 4 or 5, sensitivity 85.5% (342/400), specificity 68.3% (273/400), PPV 72.9% (342/469), and NPV 82.5% (273/331); CO-RADS 3, 4 or 5, sensitivity 91.0% (364/400), specificity

53.8% (215/400), PPV 66.3% (364/549), and NPV 85.7% (215/251). Regarding COVID-RADS, the sensitivity, specificity, PPV and NPV were as follows: COVID-RADS 3, sensitivity 65.5% (262/400), specificity 90.0% (360/400), PPV 86.8% (262/302), NPV 72.3% (360/498); COVID-RADS 2A or 3, sensitivity 69.8% (279/400), specificity 83.0% (332/400), PPV 80.4% (279/347), and NPV 73.3% (332/453); COVID-RADS 2B, 2A, or 3, sensitivity 93.0% (372/400), specificity 29.0% (116/400), PPV 56.7% (372/656), and NPV 80.6% (116/144). The RSNA expert consensus statement achieved the following sensitivity, specificity, PPV and NPV: "typical appearance", sensitivity 73.5% (294/400), specificity 82.8% (331/400), PPV 81.0% (294/363), and NPV 75.7% (331/437); and "typical appearance" or "indeterminate appearance", sensitivity 92.0% (368/400), specificity 41.0% (164/400), PPV 60.9% (368/604), and NPV 83.7% (164/196). Finally, the BSTI guidance statement marked the following sensitivity, specificity, PPV and NPV: "CLASSIC COVID-19", sensitivity 64.5% (258/400), specificity 94.0% (376/400), PPV 91.5% (258/282), and NPV 72.6 (376/518)%; "CLASSIC COVID-19" or "PROBABLE COVID-19", sensitivity 71.3% (285/400), specificity 87.3% (349/400), PPV 84.8% (285/336), and NPV 75.2% (349/464); "CLASSIC COVID-19", "PROBABLE COVID-19", or "INDTERMINATE", sensitivity 91.3% (365/400), specificity 44.8% (179/400), PPV 62.3% (365/586), and NPV 83.6% (179/214).

Interobserver variabilities of the criteria

The interobserver variabilities of COVID-19 diagnosis with 95%CI of each grading system were summarized in Table 5**.** The overall agreement was good; the average Cohen's kappa in all observers was 0.62 (95%CI: 0.58–0.66), 0.63 (95%CI: 0.58–0.68), 0.63 (95%CI: 0.57–0.69) and 0.61 (0.58-0.64) for CO-RADS, COVID-RADS, RSNA grading system, and BSTI grading system, respectively, which was not statistically significant from each other. Light's kappa was 0.56 (95%CI: 0.49-0.63) for CO-RADS, 0.55 (0.49-0.62) for COVID-RADS, 0.55 (0.49-0.62) for RSNA grading system, and 0.54 (0.48-0.62) for BSTI grading system.

Details of all 800 observations by 8 observers were summarized in Supplementary Table 1. Associations between each set of criteria were summarized in Supplementary Table 2-7.

Discussion

We conducted a comparison study of the published CT grading systems CO-RADS and COVID-RADS, and the grading system based on the RSNA expert consensus statement and the BSTI guidance statement. Although the three sets of criteria were effective in diagnosing COVID-19 with a mean AUC of about 0.80, a salient finding was that CO-RADS and the BSTI guidance statement were significantly better in distinguishing COVID-19 from non-COVID-19 etiology than COVID-RADS and the RSNA consensus statement.

All the sets of criteria were effective in terms of interobserver agreements with an average Cohen's kappa greater than 0.60. For CO-RADS, interobserver agreement was higher in the current study than in the original one (Fleiss' kappa of 0.56 vs. 0.47) [5]. For the RSNA expert consensus statement, the interobserver agreement was also higher in the current study than in a previous one (Cohen's kappa of 0.63 vs. 0.5) [10]. We attribute this to our having held a practice session using sample cases and a rigorous consensus meeting to minimize ambiguity and dependence on subjective clinical judgments and facilitate consistent image interpretation. Based on this experience we also added some remarks regarding the interpretation of these criteria to enhance clarity as summarized in Table 2, with some illustrative cases shown in Figure 2-4. As detailed in Table 2, the reasons for dividing the grading common to all the grading systems and hence candidates for possible revision were (1) the presence of more than two predominant patterns, (2) the presence of only small lesions, (3) the presence of co-existing lung disease (i.e. interstitial pneumonia or emphysema). Although some ambiguity in the terminology used may persist (i.e. (half)-rounded shape, small, homogeneous, or extensive) and a little additional modification may be required, further efforts to define the characteristics in ever greater detail may not be worthwhile considering the wide spectrum of radiological presentations and progression of COVID-19.

The diagnostic performance of CO-RADS was slightly lower than that noted in the original article (AUC of 0.84 vs. 0.91, this study vs. the CO-RADS original study [5]). This discrepancy was

attributed in part to differences in the patient cohorts studied in the original and present works. Reflecting this, as illustrated in Table 4, in this study cohort CO-RADS 1 categories showed a slightly higher RT-PCR positive rate than CO-RADS 2, while the original paper obtained the opposite result with this trend likewise observed in the other two sets of criteria). This may have resulted from differences in the inclusion criteria between the two studies; namely only patients requiring hospitalization were included in the original article, while we adopted broader inclusion criteria, namely all patients irrespective of their observation status in the present one. The difference is easy to understand since patients with COVID-19 has been reported to show different percentages of chest CT positivity in rough parallel with the severity of their symptoms: normal CT findings being observed in 46% of asymptomatic or mildly symptomatic and 21% of symptomatic patients [12]. The results of the current study confirmed that the sensitivity of chest CT is still not sufficient to use as a rule-out-tool. In contrast, the interobserver agreement was higher in the current study (Fleiss' kappa of 0.56 vs. 0.47, the current study vs. the CO-RADS original study [5]).

COVID-RADS is a relatively simple grading system based on a combination of findings with different levels of suspicion that are stratified according to their frequencies seen in COVID-19. This set of criteria, therefore, differs from the other two sets, in that it does not stratify lesions up to their axial zonal distribution (i.e., central, peripheral, or mixed distribution). However, axial zonal distribution is one of the most conspicuous and specific findings of COVID-19 as documented in previous publications [13-17]. Another limitation of this set of criteria may be related to the designation of multifocal GGO as a "typical finding." Several other diseases, including viral pneumonia of non-COVID-19 etiology, bronchial pneumonia, nonspecific interstitial pneumonia, acute interstitial pneumonitis (IP), pneumocystis jiroveci pneumonia, and drug-induced pneumonitis may show similar multifocal GGOs. In addition, discussions are needed as to whether a variety of COVID-19 and other non-COVID-19 etiologies should not best be placed in grade 2B. One example is that lesions with typical findings are downgraded to grade 2B when concomitant minor findings exist i.e. small amount of pleural effusion or emphysema, small nodules (e.g., intrapulmonary lymph nodes). However, in our experience, such co-existence is common and has

been pointed out elsewhere as well [18-20].

Compared to the other two sets of criteria based on a systematic grading system, the RSNA expert consensus statement and the BSTI guidance statement rely on groups of findings taking types of lesions, numbers, and distribution into consideration together. Based on gestalt imaging interpretation of the CT findings, these systems are easy to understand and put into practice and facilitates communication with physicians in other fields. As detailed in Table 2, one potential limitation of the RSNA consensus statement that may affects its interobserver agreements and diagnostic performance is that it does not address the presence of any co-existing lung diseases. However, as previously reported, COVID-19 pneumonia often mimics acute aggravation of IP or emphysema when superimposed on a background of IP or emphysema, thereby often causing an additional diagnostic burden (Figure 5). In contrast, the BSTI guidance statement downgrades the characteristic CT patterns ("CLASSIC COVID" or "PROBABLE COVID" classifications) when they are accompanied by other cardiopulmonary diseases (e.g. IP) to "indeterminate" classification [3]. This may explain the results of this study in which the RSNA expert consensus statement showed a higher sensitivity than the BSTI guidance statement; and in contrast, the BSTI guidance statement showed a higher specificity than the RSNA consensus statement in diagnosing COVID-19. In comparison to a previous publication on the RSNA grading system, the current study obtained comparable results with a sensitivity of 73.5% (vs. 71.6%), specificity 82.8% (vs.91.6%), and PPV 81.0% (vs. 87.8%) [10]. Previous studies also reported a similar trend observed in the current study in terms of RT-PCR positive rates for each category of the RSNA grading system with the "negative for pneumonia" classification being more frequent than "atypical" classification [9-10].

This study has various limitations. First, because of its retrospective nature, a selection bias may have been introduced. Second, the interpretation of the criteria may have been affected by our addition of supplementary remarks in an attempt to reduce ambiguity. One concern is that it may not have been consistent with the intent of the researchers who originally proposed them.

Third, the impact of this study may also have suffered from having been conducted at a single institution. However, we included eight readers with varying degrees of experience who practice at five different institutions so as to represent a broad population of readers. Fourth, only symptomatic patients were included, thereby potentially biasing the sensitivity and specificity calculated in this study. Fifth, some patients with single negative RT-PCR results were deemed negative for COVID-19. Considering the variety in the sensitivity of RT-PCR ranging from 67-98%, false negative cases were not eliminated with a single RT-PCR negative result [21]. Sixth, we did not keep the interval time constant between each session in the interpretation experiment. Ideally, the interval time should have been kept constant between each of these sessions to allow the observers to forget the details of individual cases to avoid bias, but in this study, we decided not do so to facilitate the swiftest possible publication of these findings given the pressing COVID-19 pandemic.

In conclusion, CO-RADS, COVID-RADS, the RSNA expert consensus statement, and the BSTI guidance statement provided reasonable performances and interobserver agreements in reporting the CT findings of COVID-19. Further studies will be needed to further define the clinical implications of these systems in the diagnosis of COVID-19 in a more diverse population.

Acknowledgement

The authors would like to acknowledge Ryohei Terashima for supporting statistical analysis.

References

1. Lai S, Ruktanonchai NW, Zhou L et al (2020) Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature DOI:10.1038/s41586-020-2293-x
2. Huang P, Liu T, Huang L et al (2020) Use of Chest CT in Combination with Negative RT-PCR Assay for the 2019 Novel Coronavirus but High Clinical Suspicion. Radiology 295:22-23 DOI: 10.1148/radiol.2020200330.
3. Thoracic Imaging in COVID-19 Infection, Guidance for the Reporting Radiologist (British Society of Thoracic Imaging), https://www.bsti.org.uk/media/resources/files/BSTI_COVID-19_Radiology_Guidance_version_2_16.03.20.pdf (accessed: 08/17/2020)
4. Simpson S, Kay FU, Abbara S et al (2020) Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. Radiology: Cardiothoracic Imaging 2:e200152 DOI:10.1148/ryct.2020200152
5. Prokop M, van Everdingen W, van Rees Vellinga T et al (2020) CO-RADS - A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation. Radiology 201473 DOI:10.1148/radiol.2020201473.
6. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A (2020) 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 1-13 DOI:10.1007/s00330-020-06863-0
7. Mark M Hammer, Constantine A Raptis, Travis S Henry, Amar Shah, Sanjeev Bhalla, Michael D Hope (2020) Challenges in the Interpretation and Application of Typical Imaging Features of COVID-19. Lancet Respir Med S2213-2600(20)30233-2 DOI:10.1016/S2213-2600(20)30233-2
8. Neri E, Coppola F, Larici AR, et al (2020) Structured reporting of chest CT in COVID-19 pneumonia: A consensus proposal. Insights into Imaging, 11:1-9 DOI:10.1186/s13244-020-00901-7
9. de Jaegere, TM, Krdzalic J, Fasen BA, Kwee RM. (2020) Radiological Society of North America Chest CT Classification System for Reporting COVID-19 Pneumonia: Interobserver Variability and Correlation with RT-PCR. Radiology: Cardiothoracic Imaging 2:e200213 DOI:

10.1148/ryct.2020200213

10. Ciccarese F, Coppola F, Spinelli D, et al (2020) Diagnostic Accuracy of North America Expert Consensus Statement on Reporting CT Findings in Patients with Suspected COVID-19 Infection: An Italian Single Center Experience. Radiology: Cardiothoracic Imaging 2:e200312 DOI: 10.1148/ryct.2020200312
11. Hallgren KA. (2012) Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol 8:23-34. DOI: 10.20982/tqmp.08.1.p023
12. Inui S, Fujikawa A, Jitsu M et al (2020) Chest CT Findings in Cases from the Cruise Ship "Diamond Princess" with Coronavirus Disease 2019 (COVID-19). Radiology: Cardiothoracic Imaging 2:e200110 DOI:10.1148/ryct.2020200110
13. Zhao W, Zhong Z, Xie X, Yu Q, Liu J (2020) Relation Between Chest CT Findings and Clinical Conditions of Coronavirus Disease (COVID-19) Pneumonia: A Multicenter Study. AJR Am J Roentgenol 214:1072-1077 DOI:10.2214/AJR.20.22976
14. Shi H, Han X, Jiang N et al (2020) Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infectious Diseases 20:425-434 DOI:10.1016/s1473-3099(20)30086-4
15. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A (2020) Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. AJR Am J Roentgenol 1-7 DOI:10.2214/AJR.20.23034
16. Song F, Shi N, Shan F et al (2020) Emerging 2019 Novel Coronavirus (2019-nCoV) Pneumonia. Radiology 295:210-217 DOI:10.1148/radiol.2020200274
17. Pan Y, Guan H, Zhou S et al (2020) Initial CT findings and temporal changes in patients with the novel coronavirus pneumonia (2019-nCoV): a study of 63 patients in Wuhan, China. Eur Radiol 30:3306-3309 DOI:10.1007/s00330-020-06731-x
18. Wu J, Wu X, Zeng W et al (2020) Chest CT Findings in Patients With Coronavirus Disease 2019 and Its Relationship With Clinical Features. Invest Radiol 55:257-261 DOI:10.1097/RLI.0000000000000670
19. Lomoro P, Verde F, Zerboni F et al (2020) COVID-19 pneumonia manifestations at the admission on chest ultrasound, radiographs, and CT: single-center study and comprehensive

radiologic literature review. Eur J Radiol Open 7:100231 DOI:10.1016/j.ejro.2020.100231

20. Ye Z, Zhang Y, Wang Y, Huang Z, Song B (2020) Chest CT manifestations of new coronavirus disease 2019 (COVID-19): a pictorial review. Eur Radiol 1-9 DOI:10.1007/s00330-020-06801-0

21. Kim H, Hong H, Yoon SH (2020) Diagnostic Performance of CT and Reverse Transcriptase Polymerase Chain Reaction for Coronavirus Disease 2019: A Meta-Analysis. Radiology DOI: 10.1148/radiol.2020201343

Table 1. Comparison of the CT grading systems of COVID-19

Level of suspicion	CO-RADS Category	COVID-RADS category	The RSNA expert consensus statement category	The BSTI guideline statement category
Not interpretable	**CO-RADS 0** (Scan technically insufficient for assigning a score)	Not defined.	Not defined.	Not defined.
Very low	**CO-RADS 1** (Normal or noninfectious)	**COVID-RADS 0** (Normal)	**Negative for pneumonia** (No features of pneumonia)	**NON-COVID** (70% confidence for alternative)
		COVID-RADS 1 (Atypical findings; noninfectious etiology or infectious etiology but inconsistent with COVID-19)		
Low	**CO-RADS 2** (Typical for other infection but not COVID-19)		**Atypical appearance** (Uncommonly or not reported features of COVID-19 pneumonia)	
Equivocal/unsure	**CO-RADS 3** (Features compatible with COVID-19 but also other diseases)	**COVID-RADS 2A** (Fairly typical findings) **COVID-RADS 2B** (Combination of atypical findings with typical/fairly typical findings)	**Indeterminate appearance** (Nonspecific imaging features of COVID-19 pneumonia)	**INDETERMINATE** (<70% confidence for COVID)
High	**CO-RADS 4** (Suspicious for COVID-19)			**PROBABLE COVID-19** (71-99% confidence for COVID)
Very high	**CO-RADS 5** (Typical for COVID-19)	**COVID-RADS 3** (Typical findings)	**Typical appearance** (commonly reported imaging features of greater specificity for COVID-19)	**CLASSIC COVID-19** (100% confidence for COVID)
Proven	**CO-RADS 6** (RT-PCR positive for SARS-CoV-2)	Not defined.	Not defined.	Not defined.

CO-RADS = the COVID-19 Reporting and Data System, BSTI = British Society of Thoracic Imaging.

Table 2. Summary of reasons for interobserver disagreements of the CT grading systems of COVID-19

CT grading system	Reasons for disagreements or unclear/unspecified points
CO-RADS	- Terms i.e. "(half-)rounded shape, small, homogeneous or extensive may be subjective - Categorization of GGO might be complex (addition of sample patterns as summarized below would be helpful) **[EXAMPLE] GGOs without CT findings of CO-RADS 1 or 2 in CO-RADS** Isolated GGOs (regardless of size): - Peripheral distribution, CO-RADS 4 - Otherwise, CO-RADS 3 Multifocal GGOs: - Peripheral [AND] bilateral (regardless of size): with confirmatory patterns, CO-RADS 5 without confirmatory patterns, CO-RADS 4 - Peripheral [AND] unilateral (regardless of size), CO-RADS 4 - Non-peripheral (regardless of laterality): small, CO-RADS 3 perihilar, CO-RADS 3 homogeneous [AND] extensive, CO-RADS 3 together with smooth interlobular septal thickening with or without pleural effusion, CO-RADS 3
COVID-RADS	- Does not define the following patterns: COVID-RADS 2A+3 or COVID-RADS 1+2A+3 (the former may be categorized as 3 and, the latter 2B) - Consolidation predominant pattern (late/complicated) defined in COVID-RADS 2A or 3 should be specified as that seen in organizing pneumonia (to distinguish from lobar pneumonia) - Linear opacities (late/complicated) defined in COVID-RADS 3 should be specified as those seen in organizing pneumonia (i.e. subpleural curvilinear line or perilobular pattern)
The RSNA consensus statement	- Does not address the presence of any co-existing lung disease (e.g. COVID-19 often mimics acute aggravation of IP or emphysema when superimposed on a background of IP or emphysema) - Terms e.g., "rounded/non-rounded" or "very small" may be subjective
The BSTI guideline statement	- "PROBABLE COVID-19" needs conjunctions - (and/or) [for the following reasons: (1) the three items alone may be seen in COVID-19; (2) different combination of these items denotes different likelihood in COVID-19; (3) although possible, all of these items are unlikely to be simultaneously seen in COVID-19)] - Emphysema could be included in "NON-COVID"
Common	- Judgments may be divided when two or more predominant patterns co-exist or lesions are minimal

CO-RADS = the COVID-19 Reporting and Data System, BSTI = British Society of Thoracic Imaging, GGO = ground glass opacity, IP = interstitial pneumonia.

Table 3. Characteristics of the patient cohort

Parameter	COVID-19 cases (N=50)	non-COVID-19 cases (N=50)	*P*-value
Age [years, mean ± standard deviation (range)]	59±17 (18–86)	65±24 (17–100)	.12
Gender (%)			
Men	38 (76%)	30 (60%)	.09
Women	12 (24%)	20 (40%)	–
Smoking (%)			
Yes	26 (52%)	17 (34%)	.07
No	24 (48%)	33 (66%)	–
Symptom onset to CT [days, mean (range)]	7 (3–9)	6 (2–7)	.31
Comorbidities (%)			
Absent	26 (52%)	17 (34%)	.07
Present (at least one)*	24 (48%)	33 (66%)	–
Diabetes	5 (10%)	9 (18%)	
Respiratory disease	8 (16%)	7 (14%)	
Malignancy	6 (12%)	7 (14%)	
Immune deficiency	0 (0%)	2 (4%)	
Cardiac disease	13 (26%)	18 (36%)	

* Multiple answers included

Table 4. RT-PCR positive and negative rates for each category

CO-RADS	Sum of single observations	RT-PCR positive	RT-PCR negative
CO-RADS 1	136	21 (15.4%)	115 (84.6%)
CO-RADS 2	115	15 (13.0%)	100 (87%)
CO-RADS 3	80	22 (27.5%)	58 (72.5%)
CO-RADS 4	167	84 (50.3%)	83 (49.7%)
CO-RADS 5	302	258 (85.4%)	44 (14.6%)
COVID-RADS	Sum of single observations	RT-PCR positive	RT-PCR negative
COVID-RADS 0	57	12 (21.1%)	45 (78.9%)
COVID-RADS 1	87	16 (18.4%)	71 (81.6%)
COVID-RADS 2A	45	17 (37.8%)	28 (62.2%)
COVID-RADS 2B	309	93 (30.1%)	216 (69.9%)
COVID-RADS 3	302	262 (86.8%)	40 (13.2%)
The RSNA expert consensus statement	Sum of single observations	RT-PCR positive	RT-PCR negative
Cov19Neg	109	19 (17.4%)	90 (82.6%)
Cov19Aty	87	13 (14.9%)	74 (85.1%)
Cov19Ind	241	74 (30.7%)	167 (69.3%)
Cov19Typ	363	294 (81.0%)	69 (19%)
The BSTI guidance statement	Sum of single observations	RT-PCR positive	RT-PCR negative
Non-COVID	214	35 (16.4%)	179 (83.6%)
Indeterminate	250	80 (32.0%)	170 (68.0%)
Probable COVID-19	54	27 (50.0%)	27 (50.0%)
Classic COVID-19	282	258 (91.5%)	24 (8.5%)

RT-PCR = reverse transcriptase-polymerase chain reaction, CO-RADS = the COVID-19 Reporting and Data System, Cov19Neg/Aty/Ind/Typ = negative for pneumonia/ atypical findings/ indeterminate findings/ typical findings in the RSNA expert consensus statement, BSTI = British Society of Thoracic Imaging

Table 5. Comparison of AUC and Cohen's kappa-values between the four criteria

Criteria	AUC [mean, (95%CI)]	Cohen's kappa [mean, (95%CI)]	Light's kappa [mean, (95%CI)]
CO-RADS	0.84 (0.83–0.85)	0.62 (0.58–0.66)	0.56 (0.49-0.63)
COVID-RADS	0.80 (0.78–0.81)	0.63 (0.58–0.68)	0.55 (0.49-0.62)
RSNA expert consensus statement	0.81 (0.79–0.82)	0.63 (0.57–0.69)	0.55 (0.49-0.62)
BSTI guidance statement	0.84 (0.82-0.86)	0.61 (0.58-0.64)	0.54 (0.48-0.62)
P-value for one-way repeated measures analysis of variance	< .0001*	.81	.94
P-value for paired t-test			
CO-RADS vs COVID-RADS	.0087**		
CO-RADS vs RSNA expert consensus statement	.0097**		
CO-RADS vs. BSTI guidance statement	.76		
COVID-RADS vs RSNA expert consensus statement	.39		
COVID-RADS vs. BSTI guidance statement	.0033**		
RSNA expert consensus statement vs. BSTI guidance statement	.0019**		

All AUC and Cohen's kappa value are derived from the average of each observer's value.
AUC = area under a receiver operating characteristic curves, CI = confidence interval, CO-RADS = the COVID-19 Reporting and Data System, BSTI = British Society of Thoracic Imaging, NS = not statistically significant.
* Statistically significant
** Statistically significant after Holm-Bonferroni correction

Table 6. Sensitivity, specificity, and positive and negative predictive values of the four criteria.

CO-RADS	Sensitivity	Specificity	PPV	NPV
CO-RADS 5	64.5%	89.0%	85.4%	71.5%
CO-RADS 4 or 5	85.5%	68.3%	72.9%	82.5%
CO-RADS 3, 4, or 5	91.0%	53.8%	66.3%	85.7%
COVID-RADS	Sensitivity	Specificity	PPV	NPV
COVID-RADS 3	65.5%	90.0%	86.8%	72.3%
COVID-RADS 2A or 3	69.8%	83.0%	80.4%	73.3%
COVID-RADS 2B or 2A or 3	93.0%	29.0%	56.7%	80.6%
The RSNA expert consensus statement	Sensitivity	Specificity	PPV	NPV
"Cov19Typ"	73.5%	82.8%	81.0%	75.7%
"Cov19Typ" or "Cov19Ind"	92.0%	41.0%	60.9%	83.7%
The BSTI guidance statement	Sensitivity	Specificity	PPV	NPV
"Classic COVID-19"	64.5%	92.0%	91.5%	72.6%
"Classic COVID-19" or "Probable COVID-19"	71.3%	87.3%	84.8%	75.2%
"Classic COVID-19", "Probable COVID-19", or "Indeterminate"	91.3%	44.8%	62.3%	83.6%

CO-RADS = the COVID-19 Reporting and Data System, Cov19Ind/Typ = indeterminate findings/ typical findings in the RSNA expert consensus statement, BSTI = British Society of Thoracic Imaging, PPV = positive predictive value, NPV = negative predictive value

Figures

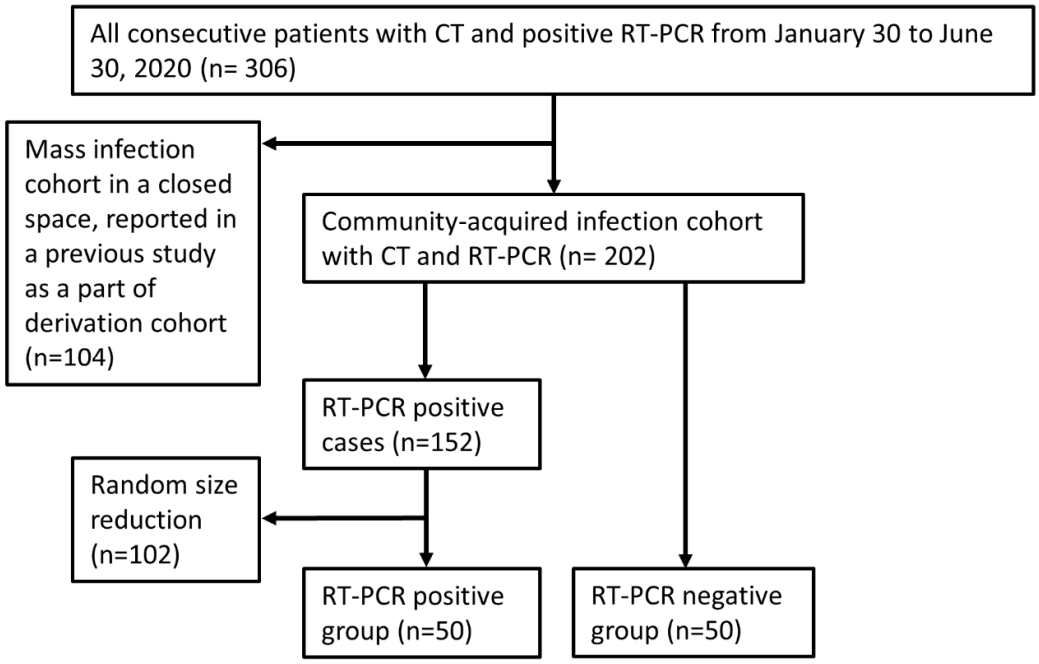
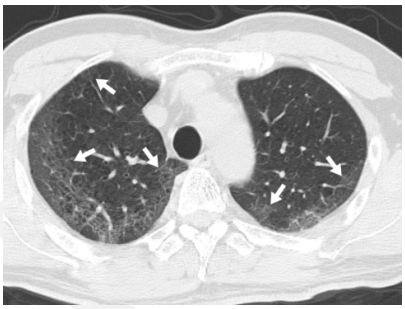
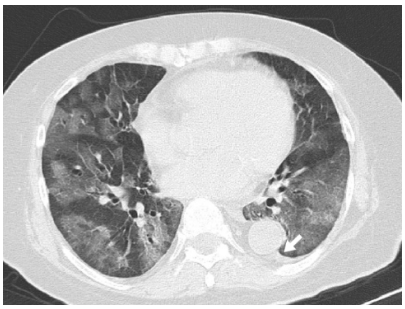


Figure 1. Flow chart illustrating the patient selection.

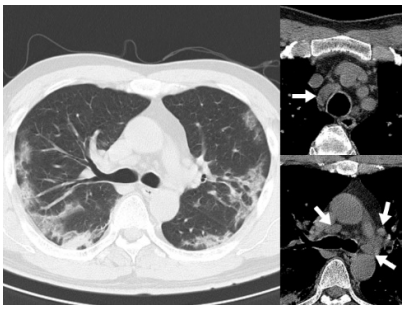
Figure 2. Cases illustrative of downgrading in COVID-RADS despite typical patterns seen in COVID-19.



(a) An axial chest CT image of a 72-year-old male with COVID-19 shows peripheral, bilateral multifocal GGOs (arrows) with a background of mild emphysema. The consensus grading was CO-RADS 5, COVID-RADS 3+1 (downgraded to 2B), typical appearance in the RSNA expert consensus statement, and classic COVID-19 in the BSTI guidance statement.

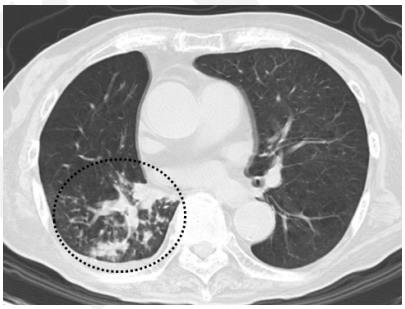


(b) An axial chest CT image of a 76-year-old female with COVID-19 shows peripheral, bilateral, multifocal GGOs with left pleural effusion (arrow). The consensus grading was CO-RADS 5, COVID-RADS 3+1 (downgraded to 2B), typical appearance in the RSNA expert consensus statement, and classic COVID-19 in the BSTI guidance statement.



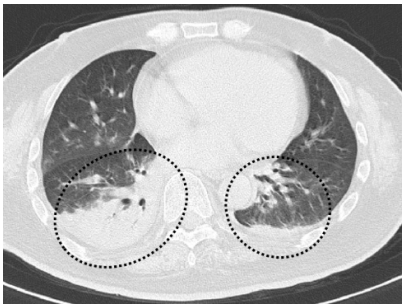
(c) An axial chest CT image of a 49-year-old male with COVID-19 shows peripheral, bilateral, multifocal GGOs with mild mediastinal lymph node enlargements (arrows). The consensus grading was CO-RADS 5, COVID-RADS 3+1 (downgraded to 2B), typical appearance in the RSNA expert consensus statement, and classic COVID-19 in the BSTI guidance statement.

Figure 3. Cases illustrative of upgrading encountered in COVID-RADS.



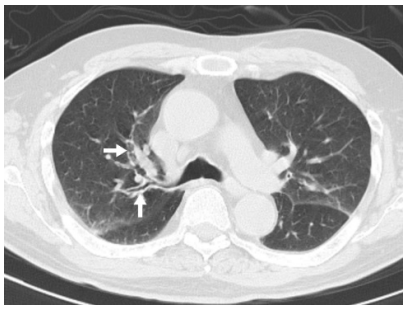
(a) An axial chest CT image of a 79-year-old female without COVID-19 shows consolidation, and centrilobular nodules without GGOs in the right lower lobe, representing typical appearances in

bronchial pneumonia (dotted circle). There was bronchial wall thickening on other images (not shown). The consensus grading was CO-RADS 2, COVID-RADS 2A+1 (=2B), atypical appearance in the RSNA expert consensus statement, and non-COVID in the BSTI guidance statement.



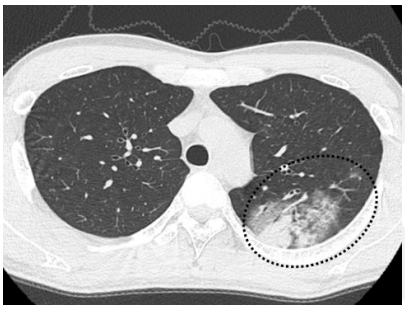
(b) An axial chest CT image of a 52-year-old male without COVID-19 shows segmental consolidation, reflecting typical appearance in lobar pneumonia (dotted circles). There was bilateral pleural effusion in this case (not shown in this figure), and the consensus grading was CO-RADS 2, COVID-RADS 2A+1 (=2B), atypical appearance in the RSNA expert consensus statement, and non-COVID in the BSTI guidance statement.

(c) An axial chest CT image of a 23-year-old male without COVID-19 shows diffuse moderate bronchial wall thickening (arrows). The consensus grading was CO-RADS 1, COVID-RADS 2A, negative in the RSNA expert consensus statement, and non-COVID in the BSTI guidance statement.

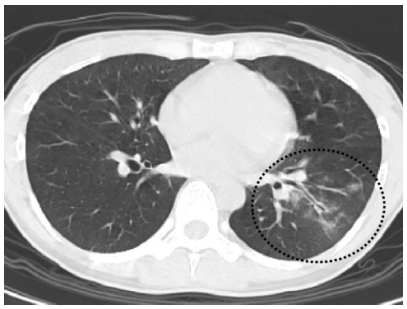


(d) An axial chest CT image of a 79-year-old male without COVID-19 shows bronchial wall thickening (arrow) but is otherwise normal. Bilateral gravity-dependent ground-glass opacities were considered non-pathological changes. The consensus grading was CO-RADS 2, COVID-RADS 2A, atypical appearance in the RSNA expert consensus statement, and non-COVID in the BSTI guidance statement. As illustrated in this case, COVID-RADS classifies bronchial thickening into higher grades than the other three sets of criteria.

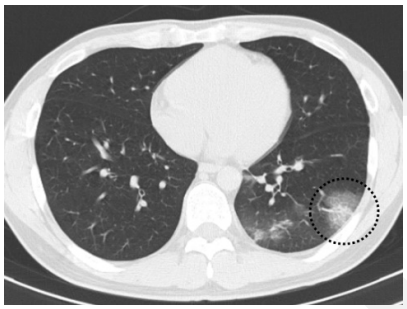
Figure 4. Cases illustrative of disaccord in gradings between each set of criteria.



(a) An axial chest CT image of a 24-year-old male without COVID-19 shows segmental GGOs and consolidation with bronchial wall thickening (dotted circle). The consensus grading was CO-RADS 2, COVID-RADS 3, indeterminate appearance in the RSNA expert consensus statement, indeterminate in the BSTI guidance statement. As illustrated in this case, if observers consider the findings compatible with typical bronchopneumonia, CO-RADS enables classification as a lower grade than do the other three sets of criteria.

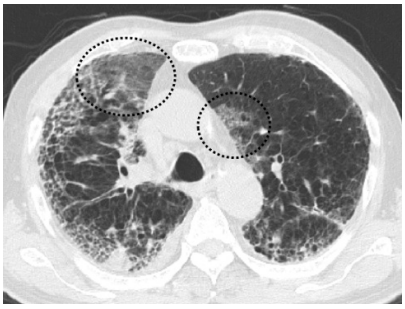


(b) An axial chest CT image of a 35-year-old female with COVID-19 shows multifocal peribronchovascular segmental centrilobular GGOs with bronchial wall thickening (dotted circle). There were centrilobular GGOs on other images (not shown). The consensus grading was CO-RADS 2, COVID-RADS 2B, atypical appearance in the RSNA expert consensus statement, and indeterminate in the BSTI guidance statement. As illustrated in this case, lesions with centrilobular distribution may be underestimated in CO-RADS and the RSNA expert consensus statement.



(c) An axial chest CT image of a 28-year-old male with COVID-19 shows peripheral, multifocal rounded GGOs with visible intralobular lines in the left lower lobe (dotted circle). The consensus grading was CO-RADS 4, COVID-RADS 3, typical appearance in the RSNA expert consensus statement, and classic COVID-19 in the BSTI guidance statement. As illustrated in this case, CO-RADS classifies lesions with unilateral distribution as grade 4 despite their typical appearance seen in COVID-19, while the other three sets of criteria assigning it to the respective highest grade.

Figure 5. Cases illustrative of diagnostic challenges with severe pre-existing pulmonary abnormalities.



(a) An axial chest CT image of a 91-year-old male without COVID-19 shows peripheral multifocal GGOs (dotted circles) in the background of usual interstitial pneumonia and moderate emphysema. There was a definite honeycomb lung destruction on other images (not shown). The consensus grading was CO-RADS 4, COVID-RADS 2B, typical appearance in the RSNA expert consensus statement, and indeterminate in the BSTI guidance statement.

(b) An axial chest CT image of a 57-year-old male with COVID-19 shows severe emphysema and airspace expansion with fibrosis. The consensus grading was CO-RADS 1, COVID-RADS 1, negative appearance in the RSNA expert consensus statement, and indeterminate in the BSTI guidance statement.

References for figure legends

8. Zhao W, Zhong Z, Xie X, Yu Q, Liu J. Relation Between Chest CT Findings and Clinical Conditions of Coronavirus Disease (COVID-19) Pneumonia: A Multicenter Study. AJR Am J Roentgenol 2020;214(5):1072-1077. doi: 10.2214/AJR.20.22976

16. Pasquier D, Le Deley MC, Tresch E, Cormier L, Duterque M, Nenan S, Lartigau E. GETUG-AFU 31: a phase I/II multicentre study evaluating the safety and efficacy of salvage stereotactic radiation in patients with intraprostatic tumour recurrence after external radiation therapy-study protocol. BMJ Open 2019;9(8):e026666. doi: 10.1136/bmjopen-2018-026666

Supplementary Tables

Supplementary Table 1. Performance and interobserver agreement of the three sets of criteria

	CO-RADS		COVID-RADS		The RSNA expert consensus statement		The BSTI guidance statement	
Observer	AUC	Cohen's kappa*	AUC	Cohen's kappa*	AUC	Cohen's kappa*	AUC	Cohen's kappa*
1	0.83 (0.74–0.90)	0.71 (0.60–0.82)	0.81 (0.72–0.88)	0.75 (0.64–0.86)	0.84 (0.75–0.90)	0.73 (0.61–0.84)	0.86 (0.78–0.92)	0.65 (0.53–0.77)
2	0.82 (0.73–0.89)	0.58 (0.45–0.70)	0.83 (0.73–0.89)	0.58 (0.45–0.71)	0.78 (0.68–0.85)	0.59 (0.46–0.72)	0.84 (0.75–0.90)	0.61 (0.53–0.77)
3	0.84 (0.75–0.91)	0.56 (0.43–0.68)	0.81 (0.73–0.88)	0.62 (0.50–0.75)	0.82 (0.73–0.88)	0.47 (0.33–0.61)	0.84 (0.75–0.90)	0.61 (0.49–0.74)
4	0.83 (0.74–0.90)	0.68 (0.57–0.79)	0.77 (0.67–0.85)	0.65 (0.52–0.77)	0.79 (0.69–0.86)	0.72 (0.60–0.83)	0.80 (0.71–0.87)	0.58 (0.45–0.71)
5	0.86 (0.78–0.92)	0.65 (0.53–0.76)	0.76 (0.66–0.83)	0.61 (0.48–0.74)	0.80 (0.70–0.87)	0.62 (0.50–0.75)	0.82 (0.72–0.88)	0.59 (0.46–0.72)
6	0.83 (0.73–0.89)	0.61 (0.49–0.73)	0.79 (0.69–0.86)	0.67 (0.55–0.79)	0.82 (0.73–0.89)	0.67 (0.55–0.80)	0.86 (0.78–0.92)	0.57 (0.44–0.70)
7	0.85 (0.75–0.91)	0.57 (0.44–0.69)	0.83 (0.65–0.91)	0.52 (0.38–0.65)	0.80 (0.71–0.87)	0.60 (0.46–0.73)	0.84 (0.75–0.90)	0.59 (0.46–0.71)
8	0.83 (0.73–0.89)	0.62 (0.50–0.74)	0.78 (0.68–0.84)	0.64 (0.51–0.76)	0.81 (0.71–0.88)	0.66 (0.54–0.78)	0.87 (0.78–0.93)	0.69 (0.57–0.80)
Overall	0.84 (0.83–0.85)	0.56 (0.54–0.58)**	0.80 (0.78–0.81)	0.55 (0.53–0.57)**	0.81 (0.79–0.82)	0.55 (0.53–0.58)**	0.84 (0.82–0.86)	0.54 (0.52–0.58)**

AUC = area under a receiver operating characteristic curves, CO-RADS = the COVID-19 Reporting and Data System, BSTI = British Society of Thoracic Imaging

Data between parentheses are 95% confidence interval.

* Cohen's kappa characteristics of classification of each observer compared to the median of the other observers.

** Fleiss' kappa

Supplementary Table 2. Association of the categories between CO-RADS and COVID-RADS

CO-RADS × COVID-RADS		CO-RADS [Sum of all single observations]				
		CO-RADS 1 n = 136 (17.0%)	CO-RADS 2 n = 115 (14.4%)	CO-RADS 3 n = 80 (10.0%)	CO-RADS 4 n = 167 (20.9%)	CO-RADS 5 n = 302 (37.8%)
COVID-RADS [Sum of all single observations]	COVID-RADS 0 n = 57 (7.1%)	57/800 (7.1%)	0/800 (0%)	0/800 (0%)	0/800 (0%)	0/800 (0%)
	COVID-RADS 1 n = 87 (10.9%)	58/800 (7.3%)	26/800 (3.2%)	3/800 (0.4%)	0/800 (0%)	0/800 (0%)
	COVID-RADS 2A n = 45 (5.6%)	7/800 (0.9%)	15/800 (1.9%)	17/800 (2.1%)	5/800 (0.6%)	1/800 (0.1%)
	COVID-RADS 2B n = 309 (38.6%)	13/800 (1.6%)	69/800 (8.6%)	56/800 (7.0%)	94/800 (11.8%)	77/800 (9.6%)
	COVID-RADS 3 n = 302 (37.8%)	1/800 (0.1%)	5/800 (0.6%)	4/800 (0.5%)	68/800 (8.5%)	224/800 (28%)

CO-RADS = the COVID-19 Reporting and Data System

Supplementary Table 3. Association of the categories between COVID-RADS and the RSNA expert consensus statement

COVID-RADS × The RSNA expert consensus statement		COVID-RADS [Sum of all single observations]				
		COVID-RADS 0 n = 57 (7.1%)	COVID-RADS 1 n = 87 (10.9%)	COVID-RADS 2A n = 45 (5.6%)	COVID-RADS 2B n = 309 (38.6%)	COVID-RADS 3 n = 302 (37.8%)
The RSNA expert consensus statement [Sum of all single observations]	Cov19Neg n = 109 (13.6%)	55/800 (6.9%)	46/800 (5.8%)	4/800 (0.5%)	4/800 (0.5%)	0/800 (0%)
	Cov19Aty n = 87 (10.9%)	1/800 (0.1%)	34/800 (4.3%)	14/800 (1.8%)	36/800 (4.5%)	2/800 (0.3%)
	Cov19Ind n = 241 (30.1%)	1/800 (0.1%)	7/800 (0.9%)	27/800 (3.4%)	155/800 (19.4%)	51/800 (6.4%)
	Cov19Typ n = 363 (45.4%)	0/800 (0%)	0/800 (0%)	0/800 (0%)	114/800 (14.3%)	249/800 (31.1%)

RADS = Reporting and Data System

Supplementary Table 4. Association of the categories between the RSNA expert consensus statement and CO-RADS

The RSNA expert consensus statement × CO-RADS		The RSNA expert consensus statement [Sum of all single observations]			
		Cov19Neg n = 109 (13.6%)	Cov19Aty n = 87 (10.9%)	Cov19Ind n = 241 (30.1%)	Cov19Typ n = 363 (45.4%)
CO-RADS [Sum of all single observations]	CO-RADS 1 n = 136 (17.0%)	101/800 (12.6%)	21/800 (2.6%)	13/800 (1.6%)	1/800 (0.1%)
	CO-RADS 2 n =115 (14.4%)	7/800 (0.9%)	58/800 (7.3%)	46/800 (5.8%)	4/800 (0.5%)
	CO-RADS 3 n = 80 (10.0%)	1/800 (0.1%)	6/800 (0.8%)	73/800 (9.1%)	0/800 (0%)
	CO-RADS 4 n = 167 (20.9%)	0/800 (0%)	2/800 (0.3%)	91/800 (11.4%)	74/800 (9.3%)
	CO-RADS 5 n = 302 (37.8%)	0/800 (0%)	0/800 (0%)	18/800 (2.3%)	284/800 (35.5%)

CO-RADS = the COVID-19 Reporting and Data System

Supplementary Table 5. Association of the categories between the BSTI guidance statement and CO-RADS

CO-RADS × The BSTI guidance statement		The BSTI guidance statement [Sum of all single observations]			
		Non-COVID n = 214 (26.8%)	Indeterminate n = 250 (31.3%)	Probable COVID-19 n = 54 (6.8%)	Classic COVID-19 n = 282 (35.3%)
CO-RADS [Sum of all single observations]	CO-RADS 1 n = 136 (17.0%)	114/800 (14.3%)	20/800 (2.5%)	1/800 (0.1%)	1/800 (0.1%)
	CO-RADS 2 n = 115 (14.4%)	66/800 (8.3%)	43/800 (5.4%)	5/800 (0.6%)	1/800 (0.1%)
	CO-RADS 3 n = 80 (10.0%)	17/800 (2.1%)	53/800 (6.6%)	8/800 (1%)	2/800 (0.3%)
	CO-RADS 4 n = 167 (20.9%)	14/800 (1.8%)	85/800 (10.6%)	21/800 (2.6%)	47/800 (5.9%)
	CO-RADS 5 n = 302 (37.8%)	3/800 (0.4%)	49/800 (6.1%)	19/800 (2.4%)	231/800 (28.9%)

CO-RADS = COVID-19 Reporting and Data System, BSTI = British Society of Thoracic Imaging system

Supplementary Table 6. Association of the categories between the BSTI guidance statement and COVID-RADS

COVID-RADS × The BSTI guidance statement		The BSTI guidance statement [Sum of all single observations]			
		Non-COVID n = 214 (26.8%)	Indeterminate n = 250 (31.3%)	Probable COVID-19 n = 54 (6.8%)	Classic COVID-19 n = 282 (35.3%)
COVID-RADS [Sum of all single observations]	COVID-RADS 0 n = 57 (7.1%)	56/800 (7%)	1/800 (0.1%)	0/800 (0%)	0/800 (0%)
	COVID-RADS 1 n = 87 (10.9%)	68/800 (8.5%)	17/800 (2.1%)	2/800 (0.3%)	0/800 (0%)
	COVID-RADS 2A n = 45 (5.6%)	23/800 (2.9%)	15/800 (1.9%)	6/800 (0.8%)	1/800 (0.1%)
	COVID-RADS 2B n = 309 (38.6%)	57/800 (7.1%)	154/800 (19.3%)	21/800 (2.6%)	77/800 (9.6%)
	COVID-RADS 3 n = 302 (37.8%)	10/800 (1.3%)	63/800 (7.9%)	25/800 (3.1%)	204/800 (25.5%)

RADS = Reporting and Data System, BSTI = British Society of Thoracic Imaging system

Supplementary Table 7. Association of the categories between the BSTI statement and the RSNA expert consensus

RSNA expert consensus statement × The BSTI guidance statement		The BSTI guidance statement [Sum of all single observations]			
		Non-COVID n = 214 (26.8%)	Indeterminate n = 250 (31.3%)	Probable COVID-19 n = 54 (6.8%)	Classic COVID-19 n = 282 (35.3%)
RSNA expert consensus statement [Sum of all single observations]	Cov19Neg n = 109 (13.6%)	103/800 (12.9%)	6/800 (0.8%)	0/800 (0%)	0/800 (0%)
	Cov19Aty n = 87 (10.9%)	55/800 (6.9%)	30/800 (3.8%)	2/800 (0.3%)	0/800 (0%)
	Cov19Ind n = 241 (30.1%)	48/800 (6%)	145/800 (18.1%)	27/800 (3.4%)	21/800 (2.6%)
	Cov19Typ n = 363 (45.4%)	8/800 (1%)	69/800 (8.6%)	25/800 (3.1%)	261/800 (32.6%)

RADS = Reporting and Data System, BSTI = British Society of Thoracic Imaging system