Implementation of the Radiological Society of North America Expert Consensus Guidelines on Reporting Chest CT Findings Related to COVID-19: A Multireader Performance Study

Authors:

Avik Som MD, PhD[1,*], Min Lang MD, MSc[1,*], Tristan Yeung[2], Denston Carey[2], Sherief Garrana[1], Dexter P. Mendoza MD[1], Efren J. Flores MD[1], Matthew D. Li MD[1], Amita Sharma, MD [1], Shaunagh McDermott, MD [1], Jo-Anne O. Shepard MD[1], Brent P. Little MD[1]

Affiliations:

[1]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

[2]Harvard Medical School, Boston, MA, USA

* Equal Effort, Co-first author

Corresponding author:

Brent Little, MD

Department of Radiology

Massachusetts General Hospital

55 Fruit Street, Boston, MA 02114

Email: blittle@partners.org

Financial and Grant Support: None

Acknowledgments: We would like to acknowledge the efforts of Nick Reid, Nicholos Joseph, and JC Panagides, of the HMS Radiology COVID research team for their aid in data collection.

Conflict of Interests and Disclosures: Dr. Little is a textbook author and editor for Elsevier and receives royalties for his prior work. Dr. Flores reports grant funding from the American College

of Radiology Innovation Fund and the National Cancer Institute Research Diversity Supplement for work not related to this manuscript.

Abbreviations: RSNA- Radiological Society of North America, CT- Computed Tomography, COVID-19 (Coronavirus Disease 2019)

Summary

In a large multireader study involving nine attendings and resident trainees, assignment of the "typical" RSNA category had strong concordance of findings across levels of training, with agreement ranging from 60% to 86%. The average sensitivity was found to be 86% (range 72%-94%), and average specificity of 80.2% (range 75%-93%) for diagnosis of COVID-19 pneumonia, and assignment of typical or indeterminate categories had an average sensitivity of 97.5% (range 94%-100%) and specificity of 54.7% (range 37%-62%); commonly reported sources of uncertainty in assignment of categories were difficulty in assessing axial distribution and the presence of two or more patterns of disease.

Key Points

1. Sensitivity and specificity of "typical appearance" for COVID-19 pneumonia on chest CT per RSNA guidelines were 86% (range 72%-94%) and 80.2% (range 75-93%), respectively.
2. There is strong concordance of findings between training levels, with agreement ranging from 60 to 86% among attendings and trainees (kappa 0.43 to 0.86).
3. Future guideline revisions should consider addressing reader uncertainty regarding assessment of axial distribution, the presence of multiple perceived patterns, and other potential sources of reader disagreements.

Abstract

Background: RSNA expert consensus guidelines provide a framework for reporting CT findings related to COVID-19, but have had limited multireader validation.

Purpose

To assess the performance of the RSNA guidelines and quantify interobserver variability in application of the guidelines in patients undergoing chest CT for suspected COVID-19 pneumonia.

Materials and Methods

A retrospective search from 1/15/20 to 3/30/20 identified 89 consecutive CT scans whose radiological report mentioned COVID-19. One positive or two negative RT-PCR tests for COVID-19 were considered the gold standard for diagnosis. Each chest CT scan was evaluated using RSNA guidelines by 9 readers (6 fellowship trained thoracic radiologists and 3 radiology resident trainees). Clinical information was obtained from the electronic medical record.

Results

There was strong concordance of findings between radiology training levels with agreement ranging from 60 to 86% among attendings and trainees (kappa 0.43 to 0.86). Sensitivity and specificity of "typical" CT findings for COVID-19 per the RSNA guidelines were on average 86% (range 72%-94%) and 80.2% (range 75-93%), respectively. Combined "typical" and "indeterminate" findings had a sensitivity of 97.5% (range 94-100%) and specificity of 54.7% (range 37-62%). A total of 163 disagreements were seen out of 801 observations (79.6% total agreement). Uncertainty in classification primarily derived from difficulty in ascertaining peripheral distribution, multiple dominant disease processes, or minimal disease.

Conclusion

The “typical appearance” category for COVID-19 CT reporting has an average sensitivity of 86% and specificity rate of 80%. There is reasonable interreader agreement and good reproducibility across various levels of experience.

Introduction

COVID-19, the disease caused by the SARS-CoV-2 virus, has become a global health emergency. Chest computed tomography (CT) has played a variety of roles in the course of the pandemic, including primary diagnosis, clinical problem solving, and assessment of potential complications. Commonly reported CT features of COVID-19 pneumonia include peripheral ground-glass opacities with or without consolidation, sometimes with an organizing lung injury appearance. (1–3) These findings are nonspecific and can be seen in a variety of infectious and non-infectious etiologies. (4–7) Early reports on the diagnostic performance of CT for detection of COVID-19 pneumonia vary substantially, with reported sensitivity ranging from 60% to 98% and specificity ranging from 25% to 53%. (8–10) The role of chest CT in screening or primary diagnosis of COVID-19 pneumonia in locales in which PCR testing is readily available is still evolving; however, chest CT plays an important role in assessing for potential complications and guide management in difficult COVID-19 cases. (11,12).

Practice patterns vary across institutions and reporting of COVID-19 has not yet been universally standardized. The Radiological Society of North America (RSNA) has recently released reporting consensus guidelines for CT findings related to COVID-19, with the goal of decreasing reporting variability, reducing uncertainty in reporting, and assisting referring providers better understand the radiological findings.(13) The guidelines contain four major categories based on the presence or absence of commonly described imaging features of COVID-19 pneumonia. Although an alternative option such as CO-RADS for categorizing CT scans have been more recently published, the RSNA consensus guidelines have been the most widely disseminated and would benefit from multi-reader validation during the implementation.(14)

The utility of implementing these reporting guidelines in radiology practices remains unclear, and the sensitivity, specificity, and inter-reader variability utilizing the four categories has not been previously studied. There is limited data on the level of inter-reader agreement, sensitivity and specificity, with early data suggesting moderate disagreement. (15) Without this empiric data, it remains uncertain for radiologists to accurately convey the level of suspicion of the CT findings and for referring clinicians to understand the clinical relevance of this information.

Indeed, understanding the diagnostic yield of a specific category may help referring clinicians understand the degree of radiologist concern for COVID-19 and the radiologist's confidence in the findings. This may influence pursuit of further diagnostic tests for COVID-19 or diagnostic workup or management for possible alternative causes of symptoms.

Imaging features of COVID-19 pneumonia are not uniform and can vary considerably, making implementation of the RSNA guidelines potentially challenging. Commonly described patterns of disease in COVID-19 are not specific to the disease and can be seen in other infections and inflammatory diseases. Patients with COVID-19 can present with negative, minimal, or atypical CT findings, or with CT findings of more than one disease. The diagnosis of COVID-19 on CT may be made by radiologists at multiple levels of training, such as the radiology resident in the emergency department, or by radiologists with varying degrees of thoracic radiology specialization. Thus, the interreader reproducibility of findings related to COVID-19 across training levels and specialization is uncertain. The purpose of this study, therefore, was to investigate the sensitivity and specificity of the RSNA/STR/ACR reporting categories for COVID-19 pneumonia and to assess interreader agreement.

Materials and Methods

Study Design and Setting

This retrospective study was performed at a large, quaternary academic medical center and associated health care system. This study was approved by the Institutional Review Board with a waiver of informed consent, and patient privacy was ensured in compliance with the Healthcare Information Portability and Accountability Act. All procedures and practices were in accordance with the Declaration of Helsinki.

Study Population

We queried our electronic imaging database for chest CT examinations performed between 1/15/2020 - 3/30/2020 and included those wherein COVID-19 pneumonia was suspected, based either on clinical indication or on radiologist suspicion as indicated in the radiology report. The gold standard for positive diagnosis of COVID-19 was at least one positive reverse transcriptase polymerase chain reaction (RT-PCR) test for COVID-19 via nasopharyngeal swab, and the gold standard for negative for COVID-19 was two consecutive negative RT-PCR results. In our healthcare system, CT scans have been used as a clinical problem-solving tool rather than for screening or primary diagnosis of COVID-19. Studies were originally ordered with either suspicion for COVID-19 despite a single negative RT-PCR (while waiting for a second test to result), concern for alternative diagnoses such as pulmonary embolism or bacterial pneumonia, or different indications such as malignancy staging with incidental findings suggestive of COVID-19 infection. Any non-diagnostic studies were excluded. 123 patients were identified by CT report or report indication including the word “COVID”, of these, 89 patients had a CT chest and a positive PCR test at any time prior to analysis, or at least two negative PCR tests and were included for analysis. In the same time period, there were 711 cases of COVID-19 that were PCR positive. In keeping with national and international guidelines, CT was used for clinical problem solving and not as a primary diagnostic modality for COVID-19 at our center, which does add selection bias for sensitivity and specificity results. 10 patients were excluded for having a CT chest and only one negative PCR test, 4 patients were excluded for having a chest CT and having no COVID PCR test. 9 patients were excluded for being identified in the pull of reports but only had a CT abdomen/pelvis available, 11 patients were excluded for missing data in the medical record or on chart review and not being available for review, leading to missing data. 1 patient of these missing data patients did turn out on repeat chart review to be COVID-19 +, but was excluded from analysis due to not being included in the original reader study. During this time period, at our institution criteria for PCR testing included symptomatology related to COVID-19 including cough, shortness of breath, fever, or being hospitalized. At the beginning of this period, due to testing restrictions, recent travel history from an endemic area or exposure to a person with known COVID-19 was required to receive a COVID-19 test. Two patients as such had a CT scan for concern for COVID-19 but were excluded for lack of a test being done. One patient had a test sent to an outside laboratory that never resulted and was excluded. One patient had a finding concerning for possible COVID-19 during staging for cholangiocarcinoma that did not fit their symptoms and the patient was set for follow-up scan in 6-8 weeks, but passed away in the interim. Clinical characteristics such as demographic data, symptoms, and comorbidities were obtained from the electronic medical record (EMR). Inability to obtain clinical data led to exclusion from the study.

CT imaging technical parameters

Protocols of CT scan varied per patient and included non-contrast chest CT, contrast-enhanced chest CT, or CT pulmonary angiography studies. CT pulmonary angiography studies had, for a subset, dual energy scans as well. All images were obtained with the patient in supine position using one of the following CT systems: Optima CT660 (GE, America), SOMATOM Drive (Siemens Healthineers, Germany), Revolution Frontier (GE, America), Lightspeed VCT (GE, America), Biograph 64 (Siemens Healthineers, Germany), SOMATOM Definition Edge (Siemens Healthineers, Germany), Discovery CT750 HD (GE, America), SOMATOM Definition Flash (Siemens Healthineers, Germany), SOMATOM Definition AS (Siemens Healthineers, Germany), SOMATOM Force (Siemens Healthineers, Germany), and Aquilion PRIME (Toshiba, Japan). The main scanning parameters were: tube voltage = 120 kVp for chest CT and for chest CT pulmonary angiography 140 kVp (plus 80 kVP for dual energy), matrix = 512 × 512, slice thickness = 1.25 mm, field of view = 440 mm × 440 mm.

Radiology Readers and Preparation

Six thoracic fellowship-trained radiology attendings (BL, SM, SG, AS, DM, EF) with 1-15 years of independent clinical practice experience, subdivided into senior (>5 years of experience BL, SM, AS, and EF), and junior attendings (<5 years of experience (DM, SG), and three radiology residents (PGY-2 to 4, AS, ML, MDL) independently reviewed all CT studies using standard PACS stations and software with standard window settings. Radiology residents were included as readers as they often provide independent preliminary reports while on call, and therefore understanding their consistency to attending reports is important. Readers were not allowed to use prior CT or follow-up CT scans to make their assessment. Each reader assigned one category from the RSNA consensus document to each study. In addition, readers reported a 0-5 score for certainty for classification of a scan into the selected RSNA category, where 5 was most certain and 0 was least certain. Reasons for uncertainty or for selection of indeterminate or atypical patterns were reported using a free text response.

Prior to CT review, radiologists studied the RSNA consensus guideline document, reviewed sample images, and had prior experience in reviewing and reporting CT examinations of patients with COVID-19 in our healthcare system. The radiology trainees were given a one-hour tutorial of the RSNA consensus guidelines with sample images from the consensus document. All radiologists were blinded to the original CT reports and to clinical diagnoses, including the PCR results for SARS-CoV-2.

RSNA Criteria

Consistent with the consensus guidelines, each examination was labeled as having “typical appearance”, “indeterminate appearance”, “atypical appearance” or “no evidence of pneumonia”. Briefly, as described in more detail in the consensus guideline publication, peripheral bilateral ground glass opacities with or without consolidation or intralobular lines, multifocal ground glass opacity with rounded morphology with or without consolidation, or reverse halo sign were assigned the category of “typical appearance”. An “indeterminate appearance” was defined as absence of typical features and presence of ground glass opacities with or without consolidation in a non-rounded, non-peripheral, perihilar or diffuse distribution, or few small ground glass opacities with a non-rounded and non-peripheral distribution. An “atypical appearance” was defined as absence of typical or indeterminate features with presence of lobar/segmental consolidation without ground glass opacities, discrete centrilobular nodules, lung cavitation, or smooth interlobular septal thickening with pleural effusion. Finally, if there were no CT findings to suggest pneumonia, it was assigned the category of “negative for pneumonia”. Examples of unanimous reader agreement for typical, indeterminate, atypical, and negative for COVID-19 pneumonia RSNA categories are shown in Figure 1.

Grading uncertainty and atypical/indeterminate findings

Reader free-text responses for reasons for uncertainty and for atypical/indeterminate findings were collated and blinded and were subsequently reviewed by 2 thoracic fellowship-trained radiologists, who by consensus discussion developed a coding scheme to capture the main themes mentioned by the readers. These 2 radiologists then sorted each response into the predetermined thematic categories determined by consensus by the 2 radiologists, blinded to reader information, clinical data, and the CT images.

Statistical Analysis

The data were analyzed with descriptive statistics, including mean and standard deviation, with categorical variables as frequencies. The kappa score, inter-rater, variable was used for statistical analysis of inter-rater agreement (the kappa scores: ≤ 0 indicates no agreement, 0.01–0.20 indicates none to slight agreement, 0.21–0.40 as fair agreement, 0.41– 0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81–1.00 as almost perfect agreement). Kappa inter-rater agreement was defined in comparison to the mode of attending responses. All trainee and attending responses were compared to the mode (majority consensus) of attending responses. Sensitivity and specificity analyses were done for each individual reader with averages calculated for attendings and trainees. Consensus reads were calculated from the mode of attending observations. Positive predictive value, negative predictive value, accuracy, and diagnostic yield of the consensus reads was compared to RT-PCR results as the gold standard. Diagnostic yield was defined as the number of SARS-CoV-2 PCR-positive patients with a CT designated as positive ("typical" or "indeterminate" RSNA categories) divided by the total number of patients in the study population. Of the individual sensitivity and specificity, averages and 95% confidence intervals were calculated. Statistical significance was set as $P < .05$. An exploratory logistic regression was completed exploring a composite clinical outcome of ICU/intubation as a function of age, gender, disagreement in RSNA classification, and RSNA grade. Statistical analysis was performed using Stata (StataCorp, College Station, TX).

Results

A total of 89 patients with CT scans meeting inclusion criteria were included in this study. The population had a mean age of 60.8 ± 16.1 years and 41(46%) were female (Table 1). Sixty-four (72%) of patients reported race as Caucasian, and the majority had presenting symptoms of fever (57%), cough (60%), and shortness of breath (58%). The most common co-morbidity was hypertension (53%). Of the patients included, 36 (40.4%) tested positive for COVID-19 by RT-PCR, and 53 (59.6%) were negative for COVID-19 infection. On average, CT scans were done 6.9 days [95% CI 5.3-8.4 days] from reported symptoms start. 84 (94.4%) of patients were admitted, with 15 (17%) requiring admission to the intensive care unit, 6 of these patients at the ICU had a positive COVID PCR test. At the time of analysis, 64 (72%) had been discharged, 22 of whom had a positive COVID PCR test, and 8 (9%) were deceased, 5 of whom had a positive COVID PCR test (Table 1). For patients who tested negative for COVID-PCR, final diagnosis included, if available, included bacterial pneumonia, viral (respiratory syncytial virus + or parainfluenza) pneumonia, sickle cell crisis, diffuse large b-cell lymphoma, heart failure exacerbation or myocardial infarction, asthma exacerbation, among others. Of the 10 patients with chest CT and only one negative PCR that were excluded from the study, 4 (40%) were male, 45 years old on average (range : 21-90), none were intubated, none went to the ICU, and the majority were diagnosed with unspecified pneumonia or viral URI. For all 34 excluded patients (with comparison to the inclusion group), the average age at time of scan was 56 years (SD 21 years, p=0.18 2-tailed t-test). 26 of the group had information on gender, of whom 13/26 (50%) were female (p=.72, Chi-squared test). 23 had ethnicity and clinical outcome results: 21/23 (91%) were Caucasian, 1/23 (4%) Hispanic, and 1/23 (4%) Asian American (p=.37, Chi-squared test). 1/23 (4%) patient was admitted to the ICU (p=.18, Fisher's exact test), and 0/23 were deceased (p=.2, Fisher's exact test). There was no significant difference between these characteristics of the included and excluded patients.

According to the majority (mode) of attending grades, 37 (41.5%) of patient CT scans were graded as "typical", 24 (27%) "indeterminate", 20 (22.5%) "atypical", and 8 (9%) "negative for pneumonia" (Table 2). Of patients in the typical group, 30 (81.1%) had a positive RT-PCR and 7 (18.9%) were negative for COVID-19 infection; in the indeterminate group, 6 (25%) were positive by RT-PCR, and 18 (75%) were negative; in the atypical group, 0 (0%) were positive by RT-PCR, and 20 (100%) were negative; in the negative for pneumonia group, 0 (0%) were positive by RT-PCR, and 8 (100%) were negative. (Table 2)

The average sensitivity and specificity of attending readers for a "typical" finding was 86% [95% CI 79.8% - 92.2%] sensitive, and 80.2% [95% CI 70.2% - 90.1%] specific. Sensitivity and specificity of typical and indeterminate grouped categorization, on average, was 97.5% [95% CI 95.1% to 99.9%], and 54.7% [95% CI 47.3% - 62%], respectively. Sensitivity and specificity of indeterminate categorization was, on average 14.2% [95% CI 7.9% to 20.5%], 69.8% [95% CI

64.7% - 74.9%], respectively. Sensitivity and specificity of atypical categorization was 2% [95% CI 0.04-4%] and 57% [95% CI 47.3% - 67.4%], respectively (Table 3). No cases classified as negative for pneumonia were associated with positive RT-PCR results.

When the mode of attending responses was considered a consensus diagnosis, sensitivity and specificity was similar to the average of different attending readers. "Typical" findings for COVID-19 on CT was 83.3% (range: 72%-94%) sensitive and 86.8% (range: 58-93%) specific for a diagnosis of COVID-19 by RT-PCR. Grouping "typical" and "indeterminate" classifications resulted in a 100% (range 94-100%) sensitivity and 53% (range 37-69%) specificity. The distribution of sensitivity and specificity was roughly the same between attendings, senior and junior, and trainees despite large differences in training (Table 3).

Among attendings, as a consensus, the positive predictive value of "typical" findings in this population was 81.1% with a negative predictive value of 88.5%. The positive predictive value was 59% for typical and indeterminate findings, 25% for indeterminate findings and 0% for atypical findings. The negative predictive value was 100% for typical and indeterminate findings, 53.8% for indeterminate findings, and 41% for atypical findings. The diagnostic yield in this retrospective study for a positive PCR among attendings as a consensus was 33.7% for "typical" findings, and 40.4% for "typical" or "indeterminate" findings.   The diagnostic accuracy in this retrospective study for a positive PCR among attendings as a consensus, was 85.4% for "typical" findings, 71.9% for typical and indeterminate findings, 46% for indeterminate findings, and 30.9% for atypical findings.

Classification of patients by reader was roughly similar between different groups (Figure 2a). A total number of 163 disagreements were seen out of 801 observations (79.6% total agreement). Using the mode (majority) of classifications from attending readers as a consensus comparison, interrater agreement among attendings was moderate to high with a kappa ranging from 0.43 to 0.86 and a range of agreement from 61% to 89% (Table 3).

There were 21 cases with at least 3 disagreements among attendings. The most common disagreement was between indeterminate and atypical categories. 11 patients (52%) had a consensus grade among attendings of indeterminate. The second most common RSNA grade for these patients was “typical” for 8 patients, and “atypical” for 3 patients. Three (14%) had an RSNA consensus grade of no pneumonia, the second most common RSNA grade for all of these patients was “atypical”. Three (14%) had a consensus grade of “atypical”, of which two had the second most common classification of “indeterminate”, and one had the classification of “no pneumonia”. Four (19%) had the consensus classification of “typical” for which the second most common classification was atypical (2) or indeterminate (2). Based on uncertainty comments the majority of indecision for these cases had issues with limited numbers of findings to make a decision, particularly leading to issues of choosing between typical and indeterminate findings.

Trainees also tended to agree moderately well with the attending modes for assigned category, with a kappa of 0.62 to 0.77 and an agreement of 74% to 84%. The primary reasons for a patient to be classified as atypical or indeterminate were a diffuse or unclear distribution, a finding of tree-in-bud or pure centrilobular nodules, focal consolidation, and pleural effusions. Less common reasons included few ground glass opacities, unilateral or central opacities, atelectasis, septal thickening, and cavitary or infarct-like lesions (Figure 2b).

An exploratory multivariable logistic regression analysis was done to assess whether a study having multiple reader disagreements was associated with a composite outcome of intubation or ICU admission. Although the results did not reach statistical significance, we found the odds ratio of having 3 disagreements (OR 0.39, p = 0.317), or 2 disagreements (OR .18, p =

0.13) to trend towards protective effects. (Supplementary Table 1). This may be because patients with disagreements tended to have less extensive pulmonary findings and thus less likely to have a negative outcome. Further study with a larger sample is necessary to evaluate this hypothesis.

Uncertainty among trainees and attendings tended to be associated with two or more dominant findings suggestive of multiple processes, minimal disease, and an ambiguous distribution or morphology of findings. Example slices from scans with significant inter-rater disagreement can be seen in Figure 3. Other sources of uncertainty included atelectasis, nodule morphology, limitations of technique, presence of pre-existing disease, or peribronchiolar pattern suggestive of organizing pneumonia (Figure 4a). The self-reported certainty score on a scale of 1 (most unsure) to 5 (most confident) tended to be between 4 and 5 on average, without a significant difference between attendings and trainees (Figure 4b). Although the absolute delta in average uncertainty is small, on average, senior attendings had less uncertainty in their classifications than trainees (p=.0011, 2-tailed t-test) or junior attendings (p=.0001, 2-tailed t-test) (Figure 4b) The average number of disagreements per case was 1.8 +/- 0.17 (Standard Error). The plurality of cases did not have any disagreements, though the majority of cases had at least one reader disagree on characterization of cases (Figure 4c). Of note, the uncertainty score for indeterminate cases was significantly reduced compared to scores for typical cases (Figure 4d).

Discussion

The RSNA consensus guidelines have provided guidelines for standardization of reporting CT findings for COVID-19 pneumonia and a framework for consistently elucidating results to referring clinicians. (13) The guidelines account for features of COVID-19 pneumonia commonly reported in the existing literature, but it is unclear how the guidelines have been interpreted and implemented among radiologists of different training levels. In this study, we assessed the diagnostic yield and diagnostic accuracy of the RSNA guidelines for CT reporting of suspected COVID-19 pneumonia, assessed interreader agreement among radiologists of different training levels for RSNA category assignment, and analyzed the distribution of RSNA consensus category scores.

Our findings concur with the literature showing the sensitivity of CT for COVID-19 pneumonia is high (10,16), as we found the combination of typical and indeterminate categorizations had an average sensitivity of 97.5% (range: 94-100%) among both attendings and radiology residents. In addition, considered together, assignment of “typical” or “indeterminate” category had a specificity of 54.7% (range: 37-60%), which matches previously reported findings, (10,16) while typical findings alone had a higher specificity at 80.2%. Selection criteria for this study attempted to replicate a tertiary center where CT is used as a problem-solving tool and not as a primary screening tool, concurrent with the RSNA and ACR guidelines. The sensitivity and specificity of the guidelines reported here must be considered with that context. While all studies categorized as negative for pneumonia were found to be RT-PCR negative in our cohort, an absence of CT findings does not exclude the possibly of COVID-19 infection. Prior studies have shown that chest CT may appear normal during early stages of infection or in those that are asymptomatic.(17) However, in these prior studies, CT was used a screening and diagnostic tool, whereas the use of CT in our study was primarily for assessment of complications or guiding management in difficult cases. Thus, there may be possible selection bias in our study as patients are all symptomatic at the time of imaging. Future studies with larger cohort size are needed to better detail the prevalence of normal CT findings in patients with COVID-19 infection.

Our results indicate a relatively high concordance among radiologists of varying level of training and experience, ranging from first year radiology residents to fellowship-trained academic radiologists with >10 years of experience, which suggests that the RSNA guidelines are clear and feasible to implement. Radiology trainees may be on the front lines of the emergency department response of COVID-19 during the day and overnight in the reading room and lately on the wards. (18) A clear guideline amenable to early trainees is more likely to result in more timely care during the pandemic. The rate of concordance was similar to slightly better to that recently reported for CO-RADS. (14)However, for a majority of cases 65 (73%), at least one reader selected a category different from the remaining readers, with 14 (15.7%) of cases having up to half the readers reporting discrepant categories. Trainees and attendings had difficulty classifying certain CT scans. Drivers of uncertainty included multiple processes, minimal disease, and an ambiguous distribution or morphology of findings. Concordant to Hickam's dictum, patients with COVID-19 may present with concurrent non-COVID-19 pathologies, with early reports suggesting 20% of patients may have additional co-infections.(7) Future guidelines should consider providing clarification in cases with CT findings from multiple categories and discuss the degree of certainty to which the categorization is placed. We also found that the terms "peripheral," "rounded opacities," and "signs of organizing pneumonia" were not well defined, causing varied interpretations for patients with peribronchovascular disease and ground glass opacities that extended centrally or diffusely. Finally, patients with limited disease remained a source of uncertainty, consistent with prior reports that minimal disease can present as atypical or confusing patterns. (19,20) This study is limited by the use of RT-PCR as a gold standard, as it has a false negative rate of up to 63% for nasopharyngeal swabs.(13,21) The biases and imperfect accuracy associated with the RT-PCR test needs to be recognized and future methods to improve the diagnostic accuracy is urgently needed. Methodologies such as composite reference standard and latent class model may be viable strategies to improve accurate detection of true COVID-19 cases.(11) This may be accomplished by combining RT-PCR results with additional test results such as chest CT and potentially identifying latent classes that are better markers for COVID-19 infection.(22) The reported sensitivity and specificity of RT-PCR testing varies across studies, with lower end estimates of 70% for sensitivity and 95% for specificity. (23,24) While there is no data on the specificity of two consecutive negative RT-PCR tests, we used two negative RT-PCR to improve the possibility that the patient did not have COVID-19, but could not exclude the possibility entirely. Until more accurate testing methods are widespread, RT-PCR remains the best validation tool available. Another limitation is that CT imaging was not correlated to timing of symptoms, with the possibility that different stages of COVID-19 infection may have a higher predilection for certain RSNA categories. Not all patients with a positive COVID-19 test receive a CT scan at our institution, which adds an element of selection bias, and that these numbers cannot be considered a study across all COVID-19 patients at our institution. The study is also limited by the single healthcare system, retrospective design. A prospective sequential inclusion of CT scans would have been ideal for studying this question, but this was not practical at our center at the start of the pandemic. In addition, because of difficulties in accessing a master list of patients at our hospital who tested positive for SARS-CoV-2 over the inclusion time period, we cannot assess the exact proportions of patients who received CT scans for suspicion of COVID-19, suspicion of PE, or had COVID-19 diagnosed at CT as an unsuspected condition. Selection bias may have resulted from our institutional use of CT as a clinical problem-solving tool rather than a method of primary COVID-19 diagnosis; in addition, application of the RSNA categories will likely be influenced by the local prevalence of disease as well as prevalence of other infectious or non-infectious etiologies, which will impact the positive and negative predictive values. A strength of our study is the relatively large number of negative COVID-PCR studies that were evaluated enabling a greater test of the RSNA guidelines. Of the 53 patients that had 2 negative RT-PCR tests, the diagnosis for these patients included bacterial pneumonia (15 patients), atypical or viral pneumonia (6 patients), cardiac related (7 related), and cancer related (7 patients); 10 patients were admitted for other unrelated reasons including trauma, cholecystitis, alcohol intoxication, sick cell crisis, bacterial colitis, liver transplant, and venous thrombosis; 8 patients did not have a definitive diagnosis. Finally, previous experience with findings of COVID-19 at chest CT can vary substantially even among radiologists with similar years of subspecialty experience, and performance of our group of readers may not reflect that of those at other institutions.

In the setting of a pandemic, the rapid implementation of standardized CT reporting has been very helpful for communicating clearly and effectively to providers about the potential of COVID-19 infection. The RSNA consensus statement serves as an important guideline for both detection of features typical for COVID-19 pneumonia and identification of features that might be seen in other infections or that might suggest alternative diagnoses.(11) In regions in which PCR testing is not severely limited, CT is useful as a tool to follow COVID-19 lung pathology and to rule out additional pathology such as PE or non-COVID pneumonia. (11) The simplicity of the RSNA consensus guidelines allow implementation by radiologists with varying levels of training. Future iterations of the guidelines should consider addressing the uncertainties found in this study to improve radiologist confidence in raising the possibility of COVID-19 pneumonia.

References

1. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. American Journal of Roentgenology. American Roentgen Ray Society; 2020;1–7.
2. Wang Y, Dong C, Hu Y, et al. Temporal Changes of CT Findings in 90 Patients with COVID-19 Pneumonia: A Longitudinal Study. Radiology. 2020;
3. Shi H, Han X, Jiang N, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. The Lancet Infectious Diseases. Lancet Publishing Group; 2020;20(4):425–434.
4. Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology. 2020;
5. Chung M, Bernheim A, Mei X, et al. CT imaging features of 2019 novel coronavirus (2019-NCoV). Radiology. Radiological Society of North America Inc.; 2020;295(1):202–207.
6. Kong W, Agarwal PP. Chest Imaging Appearance of COVID-19 Infection Case Series. Radiology: Cardiothoracic Imaging. 2020;
7. Bernheim A, Mei X, Huang M, et al. Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection. Radiology. 2020;
8. Inui S, Fujikawa A, Jitsu M, et al. Chest CT Findings in Cases from the Cruise Ship “Diamond Princess” with Coronavirus Disease 2019 (COVID-19). Radiology:Cardiothoracic Imaging. 2020;
9. Fang Y, Zhang H, Xie J, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology. 2020;
10. Ai T, Yang Z, Hou H, et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology. 2020;
11. ACR. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. 2020;
12. Rubin GD, Haramati LB, Sverzellati N, et al. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. Radiology. 2020;
13. Simpson S, Kay F, 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. Radiology: Cardiothoracic Imaging. 2020;2(2).
14. Prokop M, van Everdingen W, van Rees Vellinga T, et al. Temporal Changes of CT Findings in 90 Patients with COVID-19 Pneumonia: A Longitudinal Study. Radiology. 2020;
15. de Jaegere TM, Krdzalic J, ACM Fasen B, Kwee RM. Radiological Society of North America Chest CT Classification System for Reporting COVID-19 Pneumonia: Interobserver Variability and Correlation with RT-PCR. Radiology - Cardiothoracic Imaging. 2020;
16. Yuen Frank Wong H, Yin Sonia Lam H, Ho-Tung Fong A, et al. Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients Authors. Radiology. 2020;

17. Kim D, Quinn J, Pinsky B, Shah NH, Brown I. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. JAMA. 2020;http://www.ncbi.nlm.nih.gov/pubmed/32293646.
18. Jones J. Case Study: Answering the Call. 2020www.acr.org/.
19. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA - Journal of the American Medical Association. American Medical Association; 2020.
20. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. European Centre for Disease Prevention and Control (ECDC); 2020;25(3).
21. Zitek T. The Appropriate Use of Testing for COVID-19. The western journal of emergency medicine. 2020;http://www.ncbi.nlm.nih.gov/pubmed/32302278.
22. Fang X, Li X, Bian Y, Ji X, Lu J. Relationship between clinical types and radiological subgroups defined by latent class analysis in 2019 novel coronavirus pneumonia caused by SARS-CoV-2. European Radiology. Springer; 2020;
23. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, et al. False-negative results of initial RT-PCR assays for COVID-19: A Systematic Review. medRxiv. 2020;https://doi.org/10.1101/2020.04.16.20066787.
24. Watson J, Whiting PF, Brush JE. Interpreting a covid-19 test result. The BMJ. BMJ Publishing Group; 2020.

Table 1. Summary of Cohort Characteristics

Characteristics	**N = 89**
Mean Age ± SD, years	60.8 ± 16.1
Sex, *n* (%)	
Male	48 (53.9)
Female	41 (46.1)
Racial & Ethnic Background, *n* (%)	
Caucasian	64 (71.9)
Hispanic	12 (13.5)
Asian American	5 (5.6)
African American	4 (4.5)
Other	4 (4.5)
Presenting Symptoms, *n* (%)	
Cough	53 (59.6)
Shortness of breath	52 (58.4)
Fatigue	33 (37.1)
Chills	22 (24.7)
Chest pain	14 (15.7)
GI symptoms	12 (13.5)
Myalgia	11 (12.4)
Cognitive change	9 (10.1)
Other	7 (7.9)
Anosmia	3 (3.4)
Comorbidities, *n* (%)	
Current smoker	22 (24.7)
Current vaper	4 (4.5)
Never-smoker	43 (48.3)
Hypertension	47 (52.8)
Diabetes	28 (31.5)
Obesity	28 (31.5)
Heart failure	18 (20.2)
COPD	11 (12.4)
History of malignancy	31 (34.8)
Arrythmias	17 (19.1)
Chronic kidney disease	14 (15.7)
RT-PCR Results, *n* (%)	
Positive	36 (40.4)
Negative x2	53 (59.6)
Oxygenation Requirement, *n* (%)	
Nasal cannula/Nonrebreather	45 (51.7)
Intubation	13 (14.6)
Clinical Status, *n* (%)	
Requiring Admission	84 (94.4)
ICU	15 (16.9)
To the general medicine floor	69 (77.5)
Discharged	64 (71.9)

Deceased	8 (9.0)

SD – standard deviation, *RT-PCR* – reverse transcription polymerase chain reaction, *ICU* – intensive care unit

Table 2. Findings on Chest CT

Findings	**N = 89**	**N = 89**	**N = 89**
RSNA/STR/ACR Category, *n* (%)			
	Attendings & Trainees	*Attendings Only*	*Trainees Only*
Typical	39 (43.8)	37 (41.5)	40 (44.9)
RT-PCR Results, *n (out of,* %)			
Positive	33 (84.6)	30 (81.1)	34 (85.0)
Negative	6 (15.4)	7 (18.9)	6 (15.0)
Diagnostic Yield	37%	33.7%	38%
Indeterminate	24 (27.0)	24 (27.0)	24 (27.0)
RT-PCR Results, *n* (*out of,%)*			
Positive	3 (12.5)	6 (25.0)	2 (8.3)
Negative	21 (87.5)	18 (75.0)	22 (91.7)
Diagnostic Yield	3.4%	6.7%	2.2%
Atypical	18 (20.2)	20 (22.5)	18 (20.2)
RT-PCR Results, *n* (*out of, %)*			
Positive	0 (0.0)	0 (0.0)	0 (0.0)
Negative	18 (100.0)	20 (100.0)	18 (100.0)
Negative for pneumonia	8 (9.0)	8 (9.0)	7 (7.9)
RT-PCR Results, *n* (*out of, %)*			
Positive	0 (0.0)	0 (0.0)	0 (0.0)
Negative	8 (100.0)	8 (100.0)	7 (100.0)

RSNA – Radiological Society of North America,
STR – Society of Thoracic Radiology,
ACR – American College of Radiology

Table 3. Sensitivity/Specificity by Group: Sensitivity and specificity by attendings and trainees. Inter-reader agreement was defined by comparison of the individual reader to the mode (consensus) of attending reader scores. Average of agreement and kappa scores (with CI) in the

group of senior attending (>5 years attending experience), junior attendings (0-5 years), and trainees.

	Sensitivity(%) / Specificity(%)				*Average Inter-reader Agreement*		
N=89	Typical	Typical + Indeterminate	Indeterminate	Atypical	Agreement	Kappa	Confidence Interval [2.5%, 97.5%]
Attending average	86/80	98/55	14/70	2/57			
Senior Attending Average	83/83	96/53	15/72	3/55			
Junior Attending Average	87/92	100/58	14/66	0/62			
Trainee average	87/86	99/47	12/55	1/56			
Senior Attending 1	72/89	100/60	28/62	0/60	89%	0.84	[0.71, 0.97]
Senior Attending 2	86/83	94/58	8/75	3/57	61%	0.43	[0.31, 0.56]
Senior Attending 3	86/83	97/58	11/75	3/34	75%	0.64	[0.51, 0.77]
Senior Attending 4	86/75	94/37	11/75	6/70	74%	0.62	[0.48, 0.75]
Junior Attending 5	92/93	100/62	8/70	0/57	81%	0.73	[0.60, 0.86]
Junior Attending 6	81/91	100/53	19/62	0/66	90%	0.86	[0.73, 0.98]
Trainee 1	78/87	97/58	19/72	3/57	76%	0.	[0.53, 0.79]

						66	
Trainee 2	97/87	100/38	3/51	0/70	74%	0.62	[0.49, 0.75]
Trainee 3	86/83	100/45	14/43	0/40	84%	0.77	[0.64, 0.90]

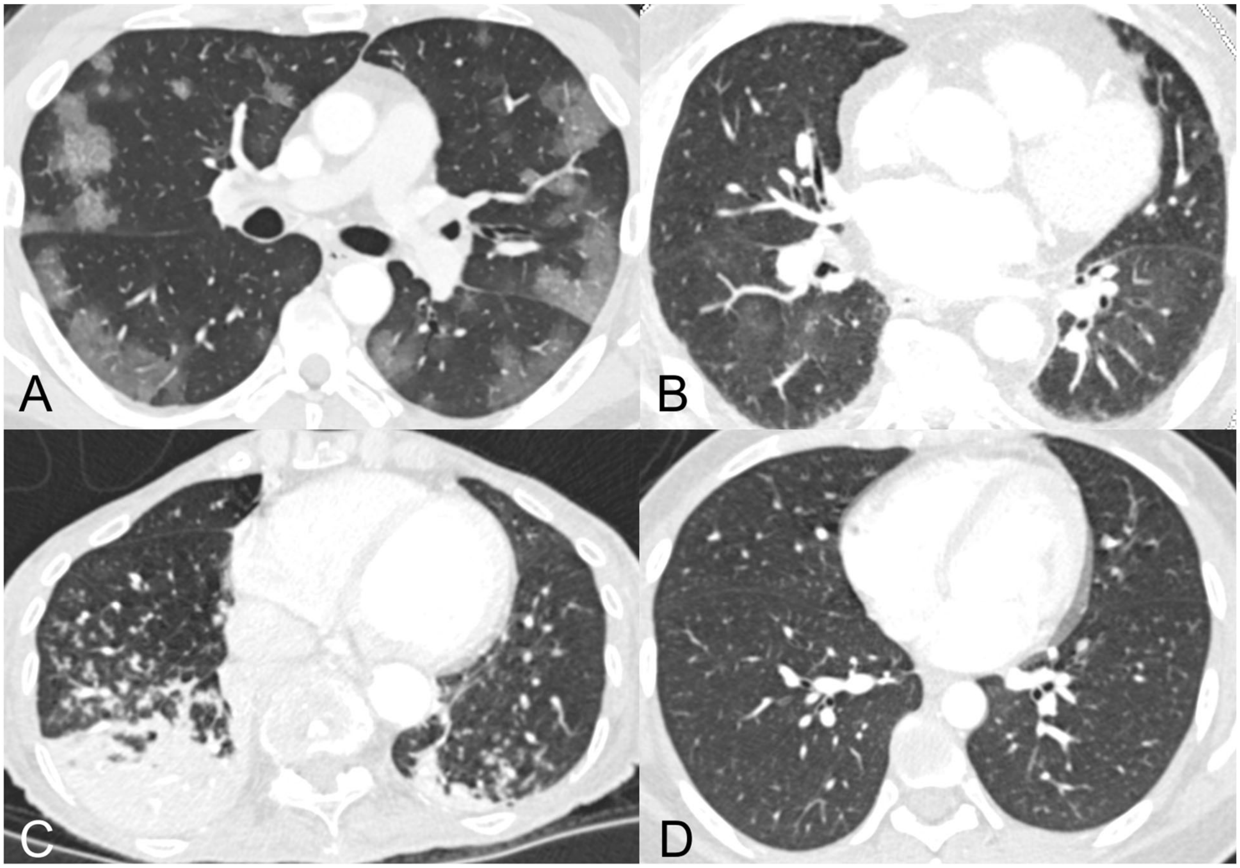

Figure 1. Examples of cases assigned the same RSNA consensus COVID-19 category by all readers.
A) “Typical” category assigned to CT of 53-year-old man with COVID-19 pneumonia who presented after 2 weeks of cough, congestion, and fevers. Axial CT image shows multiple ground glass opacities with a peripheral predominance bilaterally, many with a round morphology.
B) “Indeterminate” category assigned to CT of 82-year-old woman who presented with fever, exertional dyspnea, palpitations, and chest pain, with 2 PCRs negative for SARS-CoV-2. Axial CT image shows a small amount of ground glass opacity with a central predominance in the perihilar regions bilaterally.
C) “Atypical” category assigned to CT of a 79-year-old woman who presented with fever, productive cough, dyspnea, and hypoxemia; 2 PCRs were negative for SARS-CoV-2. Axial CT image shows tree-in-bud nodules and consolidation in the lower lobes bilaterally, a pattern suggesting aspiration/pneumonia.
D) “Negative for pneumonia” category assigned to CT of a 30-year-old woman who presented with one week of dry cough, sore throat, and severe fatigue; 2 PCRs were negative for SARS-CoV-2. Axial CT image shows a normal appearance of the lungs. Final diagnosis of symptoms was attributed to recurrent rheumatic myopericarditis within the context of her history of juvenile rheumatoid arthritis.

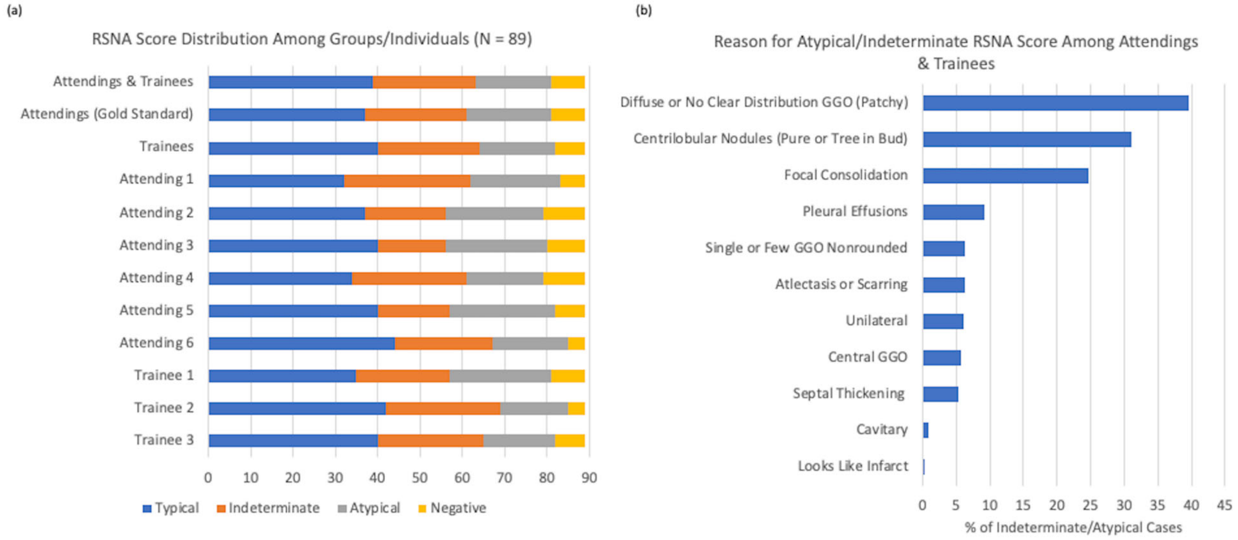

Figure 2. Reasoning for Atypical/Indeterminate RSNA Score Among Attendings & Trainees. a) Distribution of scores among different readers. b) Percentage of cases with particular reasons for being assigned a category of indeterminate or atypical.

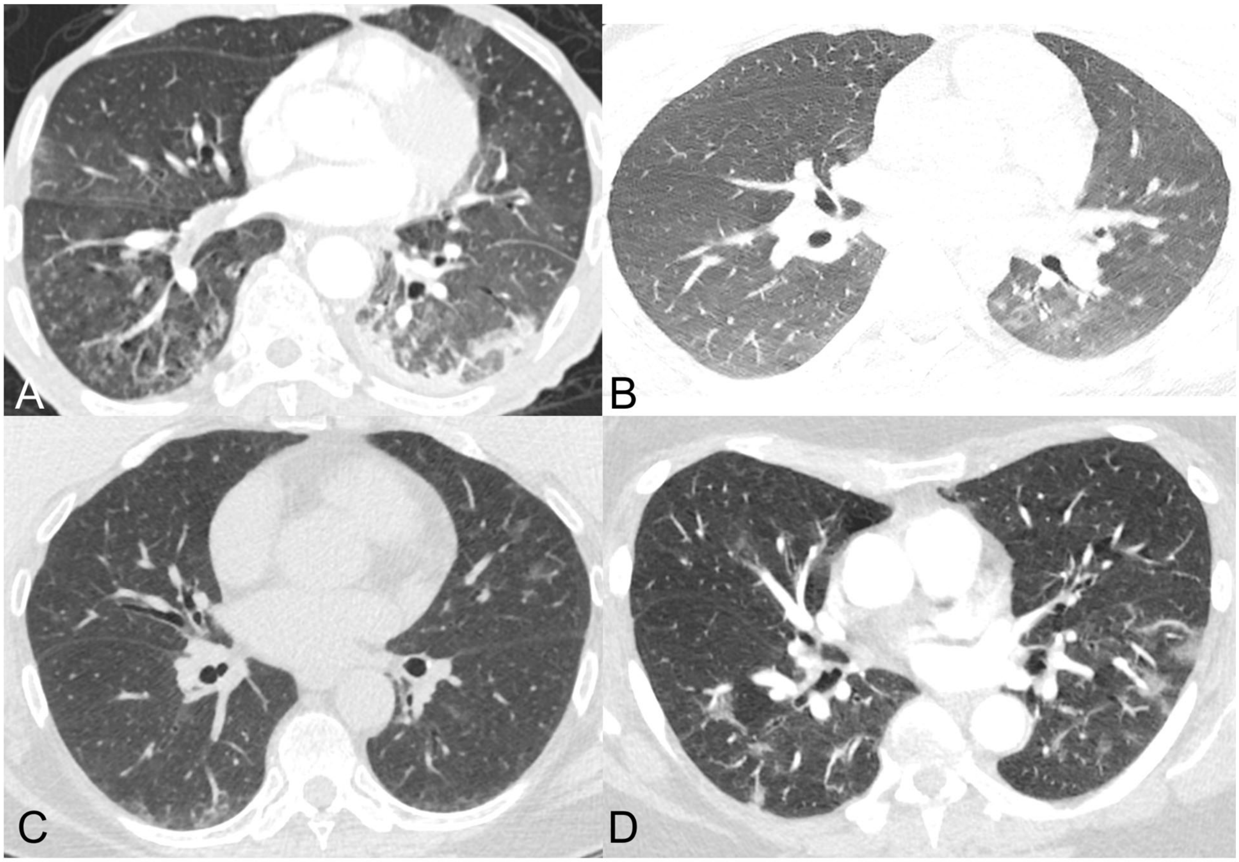

Figure 3. Examples of cases for which there was significant disagreement in assignment of RSNA consensus COVID-19 category.

A) 67-year-old man with clinical signs of pneumonia and 4 negative PCRs for COVID-19 with sputum samples positive for *streptococcus pneumoniae*. Axial CT image shows a combination of tree-in-bud centrilobular nodules in the lower lobes and peripheral ground glass opacity and consolidation in the left lower lobe. Categories 3, 2, and 1 were assigned by 4, 3, and 2 readers respectively.

B) 23-year-old man with 2 negative PCR results for SARS-CoV-2 and presumed aspiration or non-COVID-19 infection. Axial CT image shows minimal patchy ground glass opacities in the left lower lobe; there was a question of atelectasis or subtle peripheral ground glass opacity in the posterior right lower lobe. Categories 3, 2, 1, and 0 were assigned by 1, 6, 1, and 1 readers respectively.

C) 64-year-old woman with PCR-proven COVID-19 pneumonia who presented with fever, productive cough, fatigue, and anosmia. Axial CT image shows patchy ground glass opacities in the lingula and a small amount of peripheral ground glass opacity and atelectasis in the posterior lower lobes. Categories 3, 2, 1, and 0 were assigned by 4, 3, 1, and 1 readers respectively. Reasons given by readers for uncertainty included doubts about peripheral distribution, and difficulty in classification in the setting of minimal disease and posterior atelectasis.

D) 65-year-old woman with PCR-proven COVID-19 pneumonia who presented with palpitations, back pain, and low-grade fevers. Axial CT image shows patchy ground glass opacities bilaterally. Categories 3 and 2 were assigned by 5 and 4 readers respectively. Reasons given by readers for uncertainty included difficulty in classifying as peripheral or diffuse and questionable morphology of the ground glass opacities.

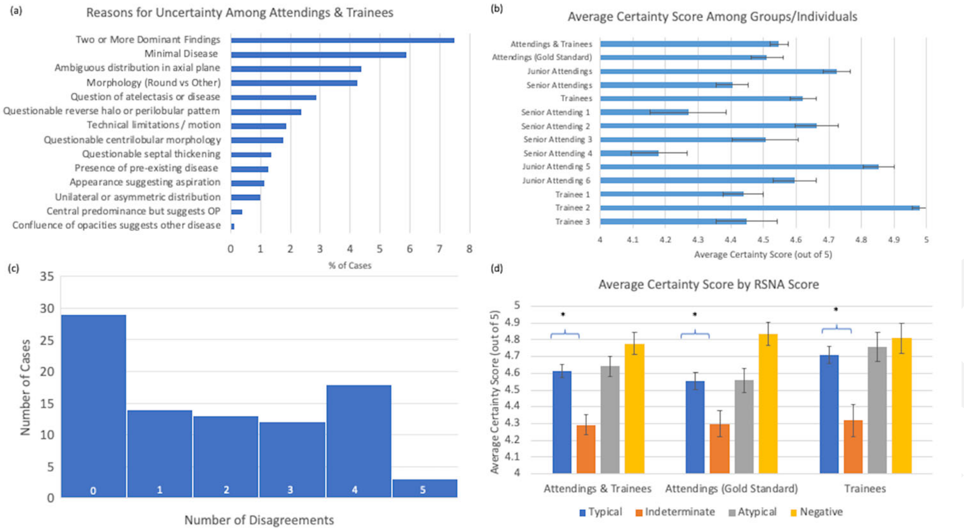

Figure 4. Uncertainty Among Attending & Trainees: a) Reasons for uncertainty among cases as a percentage of all cases reviewed. OP- organizing pneumonia b) Average certainty scores between attendings and trainees. c) Histogram of number of readers with scores discrepant from attending consensus. d) Average certainty score by RSNA categorization, * indicates statistical significance,- p<.05 (2-tailed t-test)

Supplementary Table 1: Exploratory multivariable logistic regression for effect of disagreement on composite ICU/Intubation risk as a function of age, gender, disagreement, or RSNA grade.

Intubated/ICU Composite Outcome	*Odds Ratio*	*Std. Err.*	*P-value*	*95% CI*
Age	1.02	0.02	0.29	(0.98,1.06)
Female	1.01	0.59	0.98	(0.32, 3.15)
Disagreement (# of attendings disagreeing)				
1	1.6	1.09	0.53	(0.39,6.17)
2	0.18	0.20	0.13	(0.02,1.65)
3	0.39	0.37	0.32	(0.06,2.43)
4	1	-	-	-
RSNA Grade (Attending Consensus)				
No Pneumonia	1	-	-	-
Typical	1.11	0.81	0.89	(0.27,4.61)
Indeterminate	0.51	0.42	0.41	(0.10,2.55)
Atypical	1	-	-	-