Interobserver and Intraobserver Variability in the CT Assessment of COVID-19 Based on RSNA Consensus Classification Categories

Mohamad Hadied  
*Henry Ford Health, MHADIED1@hfhs.org*

Parth Y. Patel  
*Henry Ford Health, PPatel31@hfhs.org*

Peter Cormier  
*Henry Ford Health, pcormie1@hfhs.org*

Neo Poyiadji  
*Henry Ford Health, npoyiad1@hfhs.org*

Mariam Salman  
*Henry Ford Health, msalman1@hfhs.org*

*See next page for additional authors*

Follow this and additional works at: [https://scholarlycommons.henryford.com/radiology_articles](https://scholarlycommons.henryford.com/radiology_articles)

**Recommended Citation**


This Article is brought to you for free and open access by the Diagnostic Radiology at Henry Ford Health Scholarly Commons. It has been accepted for inclusion in Diagnostic Radiology Articles by an authorized administrator of Henry Ford Health Scholarly Commons.
Interobserver and Intraobserver Variability in the CT Assessment of COVID-19 Based on RSNA Consensus Classification Categories

Mohamad O. Hadied, MD, Parth Y. Patel, MD, Peter Cormier, MD, Neo Poyiadji, MD, Mariam Salman, MD, Chad Klochko, MD, Jeffrey Nadig, MD, Thomas Song, MD, Ed Peterson, PhD, Nick Reeser, MD

Purpose: To assess the interobserver and intraobserver agreement of fellowship trained chest radiologists, nonchest fellowship-trained radiologists, and fifth-year radiology residents for COVID-19-related imaging findings based on the consensus statement released by the Radiological Society of North America (RSNA).

Methods: A survey of 70 chest CTs of polymerase chain reaction (PCR)-confirmed COVID-19 positive and COVID-19 negative patients was distributed to three groups of participating radiologists: five fellowship-trained chest radiologists, five nonchest fellowship-trained radiologists, and five fifth-year radiology residents. The survey asked participants to broadly classify the findings of each chest CT into one of the four RSNA COVID-19 imaging categories, then select which imaging features led to their categorization. A 1-week washout period followed by a second survey comprised of randomly selected exams from the initial survey was given to the participating radiologists.

Results: There was moderate overall interobserver agreement in each group (κ coefficient range 0.45-0.52 ± 0.02). There was substantial overall intraobserver agreement across the chest and nonchest groups (κ coefficient range 0.61-0.67 ± 0.06) and moderate overall intraobserver agreement within the resident group (κ coefficient 0.58 ± 0.06). For the image features that led to categorization, there were varied levels of agreement in the interobserver and intraobserver components that ranged from fair to perfect kappa values. When assessing agreement with PCR-confirmed COVID status as the key, we observed moderate overall agreement within each group.

Conclusion: Our results support the reliability of the RSNA consensus classification system for COVID-19-related image findings.

Key Words: Interobserver variability; intraobserver variability; RSNA COVID-19 chest CT consensus classification categories.

INTRODUCTION

In response to the COVID-19 pandemic, the Radiological Society of North America (RSNA) released a consensus statement on the reporting of chest CT findings related to COVID-19 in March of 2020 (1). To date, there have been over 21.2 million confirmed cases worldwide, with over 2 million confirmed cases in the United States (2,3). Most professional radiological societies, as well as the Center for Disease Control, have recommended against the use of screening chest CTs for the detection of COVID-19. Despite this, chest CTs are performed frequently on patients with suspected or confirmed COVID-19 (4-10). To best equip clinicians with the necessary tools to make informed decisions with regards to the management of COVID-19, there exists a need for accurate and precise reporting of imaging findings. The RSNA consensus statement proposed four categories for standardized CT reporting based on expert consensus with endorsement from the Society of Thoracic Radiology and the American College of Radiology (1). The purpose of this study is to assess the interobserver and intraobserver variability of COVID-19-related imaging findings based on the criteria outlined by the consensus statement released by RSNA.

METHODS

Sample Selection

Institutional review board approval was obtained for this study and patients’ consents were waived. An automated data
pull of all chest CT exams performed on patients from March 16, 2020 to April 18, 2020 at a single institution yielded 1500 eligible patients (Fig 1). All 1500 electronic record charts were manually reviewed to exclude those who were not tested for COVID-19. This process yielded 893 records. These records were then manually filtered for mention of the RSNA classification system within the final CT dictation to insure a diversity of cases, which yielded a total of 210 records. Of these 210 exams, 122 displayed findings that were reported as “Typical,” 37 of the exams were determined to be “Indeterminate,” 17 were deemed “Atypical” and 34 were reported “Negative” based on the RSNA classification system. Manual search of the electronic medical record was done to determine the COVID-19 status of the 210 patients, which was tested for at this institution via nasal swab polymerase chain reaction (PCR).

**Survey Design and Statistical Analysis**

A statistical analysis demonstrated 70 of the 210 exams would be necessary to appropriately power the study (11). Seventy exams were randomly selected from the 210 available chest CTs. The

![Flow diagram of patient exam selection.](image-url)
70 exams were randomized, deidentified, and incorporated into an online research picture archiving and communication system developed at our institution. Survey participants were blinded to the image selection process, as well as the initial chest CT classification, and were told they would be receiving a random set of chest CTs from the recent months to evaluate utilizing the RSNA classification system. Of the 70 included chest CTs, COVID-19 positive and COVID-19 negative patients were included to a ratio of 46:24.

The survey asked participants to broadly classify the findings of each chest CT into one of the four RSNA COVID-19 imaging categories (Figs 2 and 3). After category selection, participants were asked to specify which images features led to their classification choice (Figs 2 and 3).

Figure 2. Consensus RSNA classification system for chest CT imaging findings related to COVID-19 with four categories and suggested reporting language.
Three different groups participated in the survey, varying in both level of training and subspecialization. The measure of inter-rater agreement for the categorical measurement of COVID-19 was the Kappa statistic. We followed Fleiss, Levin, and Paik’s technique for multiple ratings per scan with different raters (12). Statistical analysis determined that a minimum of four participants per group was necessary to give the study a power of 90% to detect a Kappa statistic of 0.20 or more. The study participants included five fellowship-trained chest radiologists (Group 1), five nonchest fellowship-trained radiologists consisting of two emergency and three abdominal-trained radiologists (Group 2), and five fifth-year radiology residents (Group 3). Survey participants were blinded to both the COVID-19 status of each patient, as well as the original radiology report.

The intraobserver survey component was completed after 1 week of washout time and contained 24 exams randomly selected from the initial survey for a power of 80% to detect a Kappa statistic of 0.40 or more. The image order and deidentified accession numbers were randomized again for the intraobserver survey. The 1-week washout was determined to be sufficient to prevent a learning curve given the randomization of image selection. COVID-19 positive and COVID-19 negative patients were included in the intraobserver survey to a ratio of 9:15.

**RESULTS**

The average year of experience of the five chest fellowship-trained radiologists was 13.2 years with a range of 3-35 years. The average year of experience of the nonchest fellowship-trained radiologists was 17.2 years with a range of 2-27 years. Five tables display the kappa results as discussed below, only statistically significant k values were included within the tables.

**Interobserver Agreement**

There was moderate overall agreement ($\kappa$ coefficient range 0.45-0.52 ± 0.02) for the four RSNA categories across all three participating groups (Table 1). For each individual category, there was moderate agreement across all three groups for the “Typical” category ($\kappa$ coefficient range 0.44-0.55 ± 0.02), fair agreement across all groups for the “Indeterminate” category ($\kappa$ coefficient range 0.20-0.36 ± 0.01-0.06), fair to moderate overall agreement for the “Atypical” category ($\kappa$ coefficient range 0.37-0.45 ± 0.02), and substantial...
agreement across all three groups for the “Negative” category ($\kappa$ coefficient range 0.71-0.78 ± 0.02).

For the secondary questions within each category as defined on Table 2, there was fair-moderate overall agreement across all three groups for the first set of “Typical” imaging features (i.e., peripheral, bilateral, ground-glass opacities [GGO] with or without consolidation or visible intralobular lines). There was substantial to perfect agreement among participating chest and nonchest radiologists for the image features of the “Indeterminate” category. The remainder of the agreement for the secondary questions varied from zero to perfect agreement as described on Table 2.

**Intraobserver Agreement**

There was substantial overall intraobserver agreement across the chest and nonchest groups ($\kappa$ coefficient range 0.61-0.67 ± 0.05-0.06) and moderate overall agreement within the resident group ($\kappa$ coefficient 0.58 ± 0.06). For the individual categories, there was moderate intraobserver agreement across all three groups for the “Typical” and “Atypical” categories, with fair to moderate in the “Indeterminate” category, and almost perfect in the “Negative” category (Table 3).

**Correlation With PCR-Confirmed COVID-19 Status**

Using PCR-confirmed COVID-19 status (i.e., positive vs negative) as the standard for the patient exams included in our study, there was moderate overall agreement across all three groups ($\kappa$ coefficient range 0.53-0.57 ± 0.05, Table 5). Results were best when agreed upon “Typical” and “Indeterminate” categories were combined and corresponded with moderate overall agreement to COVID-19 positive status. The COVID-19 negative group was a combination of moderately agreed upon “Atypical” and “Negative” categories.

**TABLE 1. Kappa Coefficients With Standard Errors for the Interobserver Agreement Survey**

<table>
<thead>
<tr>
<th>Group</th>
<th>Overall</th>
<th>Typical</th>
<th>Indeterminate</th>
<th>Atypical</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest</td>
<td>0.52 ± 0.02</td>
<td>0.51 ± 0.02</td>
<td>0.36 ± 0.02</td>
<td>0.43 ± 0.02</td>
<td>0.78 ± 0.02</td>
</tr>
<tr>
<td>Nonchest</td>
<td>0.49 ± 0.02</td>
<td>0.55 ± 0.02</td>
<td>0.32 ± 0.02</td>
<td>0.37 ± 0.02</td>
<td>0.71 ± 0.02</td>
</tr>
<tr>
<td>Resident</td>
<td>0.45 ± 0.02</td>
<td>0.44 ± 0.02</td>
<td>0.20 ± 0.06</td>
<td>0.45 ± 0.02</td>
<td>0.76 ± 0.02</td>
</tr>
</tbody>
</table>

**TABLE 2. Kappa Coefficients for Secondary Questions of the Interobserver Survey**

<table>
<thead>
<tr>
<th>Imaging Features</th>
<th>Chest</th>
<th>Nonchest</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peripheral, bilateral, GGO with or without consolidation or visible intralobular lines (i.e., crazy paving)</td>
<td>0.57 ± 0.19</td>
<td>0.37 ± 0.10</td>
<td>0.15 ± 0.17</td>
</tr>
<tr>
<td>Multifocal GGO of rounded morphology with or without consolidation or visible intralobular lines (“crazy paving”)</td>
<td>0.30 ± 0.12</td>
<td>0.29 ± 0.16</td>
<td>0.24 ± 0.14</td>
</tr>
<tr>
<td>Reverse halo sign or other findings of organizing pneumonia</td>
<td>-0.03 ± 0.16</td>
<td>0.08 ± 0.23</td>
<td>-0.15 ± 0.32</td>
</tr>
<tr>
<td>Indeterminate:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multifocal, diffuse, perihilar, or unilateral GGO with or without consolidation lacking a specific distribution and are nonrounded or nonperipheral</td>
<td>1.00 ± 0.30</td>
<td>0.73 ± 0.26</td>
<td>0.19 ± 0.23</td>
</tr>
<tr>
<td>Few very small GGO with a nonrounded and nonperipheral distribution</td>
<td>1.00 ± 0.28</td>
<td>0.73 ± 0.26</td>
<td>0.31 ± 0.20</td>
</tr>
<tr>
<td>Atypical:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isolated lobar or segmental consolidation without GGO</td>
<td>0.65 ± 0.23</td>
<td>0.47 ± 0.19</td>
<td>0.60 ± 0.18</td>
</tr>
<tr>
<td>Discrete small nodules (centrilobular, tree-in-bud)</td>
<td>0.38 ± 0.36</td>
<td>0.47 ± 0.14</td>
<td>0.61 ± 0.17</td>
</tr>
<tr>
<td>Lung cavitation</td>
<td>-0.37 ± 0.66</td>
<td>1.00 ± 0.78</td>
<td>0.16 ± 0.58</td>
</tr>
<tr>
<td>Smooth interlobular septal thickening with pleural effusion</td>
<td>1.00 ± 0.52</td>
<td>0.39 ± 0.29</td>
<td>0.70 ± 0.20</td>
</tr>
</tbody>
</table>

**TABLE 3. Kappa Coefficient With Standard Errors for Intraobserver Variability Survey**

<table>
<thead>
<tr>
<th>Group</th>
<th>Overall</th>
<th>Typical</th>
<th>Indeterminate</th>
<th>Atypical</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest</td>
<td>0.67 ± 0.05</td>
<td>0.57 ± 0.08</td>
<td>0.58 ± 0.08</td>
<td>0.50 ± 0.16</td>
<td>0.92 ± 0.04</td>
</tr>
<tr>
<td>Nonchest</td>
<td>0.61 ± 0.06</td>
<td>0.53 ± 0.09</td>
<td>0.48 ± 0.08</td>
<td>0.45 ± 0.12</td>
<td>0.92 ± 0.04</td>
</tr>
<tr>
<td>Resident</td>
<td>0.58 ± 0.06</td>
<td>0.60 ± 0.09</td>
<td>0.38 ± 0.09</td>
<td>0.46 ± 0.11</td>
<td>0.88 ± 0.05</td>
</tr>
</tbody>
</table>
Below a few case examples of chest CT exams included in our study. The first two of which (Figs 4 and 5) are cases that received unanimous agreement across all 15 participants as “Typical” and “Atypical” respectively for imaging findings related to COVID-19 based on the RSNA consensus reporting guidelines. The next two (Figs 6 and 7) are cases that received poor observer agreement as described.

**TABLE 4. Kappa Coefficient for Secondary Questions of the Intraobserver Variability Survey**

<table>
<thead>
<tr>
<th>Secondary Imaging Features</th>
<th>Chest</th>
<th>Nonchest</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peripheral, bilateral, GGO with or without consolidation or visible intralobular lines (i.e., crazy paving)</td>
<td>0.55 ± 0.08</td>
<td>0.41 ± 0.10</td>
<td>0.33 ± 0.11</td>
</tr>
<tr>
<td>Multifocal GGO of rounded morphology with or without consolidation or visible intralobular lines (“crazy paving”)</td>
<td>0.50 ± 0.11</td>
<td>0.36 ± 0.13</td>
<td>0.50 ± 0.11</td>
</tr>
<tr>
<td>Reverse halo sign or other findings of organizing pneumonia</td>
<td>0.76 ± 0.13</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Indeterminate:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multifocal, diffuse, perihilar, or unilateral GGO with or without consolidation lacking a specific distribution and are nonrounded or nonperipheral</td>
<td>0.61 ± 0.07</td>
<td>0.52 ± 0.08</td>
<td>0.41 ± 0.09</td>
</tr>
<tr>
<td>Few very small GGO with a nonrounded and nonperipheral distribution</td>
<td>–0.01 ± 0.01</td>
<td>0.66 ± 0.32</td>
<td>0.11 ± 0.14</td>
</tr>
<tr>
<td>Atypical:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isolated lobar or segmental consolidation without GGO</td>
<td>0.27 ± 0.23</td>
<td>0.46 ± 0.14</td>
<td>0.57 ± 0.13</td>
</tr>
<tr>
<td>Discrete small nodules (centrilobular, tree-in-bud)</td>
<td>0.70 ± 0.14</td>
<td>0.72 ± 0.12</td>
<td>0.47 ± 0.16</td>
</tr>
<tr>
<td>Lung cavitation</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Smooth interlobular septal thickening with pleural effusion</td>
<td>–</td>
<td>0.49 ± 0.30</td>
<td>0.65 ± 0.19</td>
</tr>
</tbody>
</table>

**TABLE 5. Kappa Coefficient Using PCR-Confirmed COVID-19 Status as Key**

<table>
<thead>
<tr>
<th>Group</th>
<th>Using COVID-19 Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest</td>
<td>0.55 ± 0.05</td>
</tr>
<tr>
<td>Nonchest</td>
<td>0.53 ± 0.05</td>
</tr>
<tr>
<td>Resident</td>
<td>0.57 ± 0.05</td>
</tr>
</tbody>
</table>

**Chest CT Cases**

Below a few case examples of chest CT exams included in our study. The first two of which (Figs 4 and 5) are cases that received unanimous agreement across all 15 participants as “Typical” and “Atypical” respectively for imaging findings related to COVID-19 based on the RSNA consensus reporting guidelines. The next two (Figs 6 and 7) are cases that received poor observer agreement as described.

**DISCUSSION**

Professional radiological societies and the Center for Disease Control currently recommend against the use of chest CTs for the detection of COVID-19. Despite this, clinicians frequently order these imaging studies to guide clinical management of suspected or confirmed COVID-19 patients (4-10). The role of chest CTs in the management of COVID-19 patients is evolving, especially given recent literature describing increased rates of thromboembolic complications such as increased rates of pulmonary embolism among COVID-19 patients (14). Due to the frequent incorporation of chest CTs for the detection of COVID-19 and the management of active COVID-19 cases...
The presence of both features in a CT, posed difficulty in determining whether to categorize CT findings as “Typical” or “Indeterminate.” This would explain why participants had some difficulty in determining whether to categorize CT findings as “Indeterminate” category or the “Typical” category. The potential solution to this dilemma would be emphasizing the importance of the RSNA classification system prior to our survey other than their own acquired experience with reading chest CTs during the peak of COVID-19 cases. This is a potential limitation of the study as the lack of familiarity with the RSNA classification system may lead to misinterpreting or missing COVID-19 specific imaging findings, this would be expected to affect nonchest-trained participants to a greater degree.

When correlating the interobserver agreement results to the PCR-confirmed COVID-19 status of the patients, this study demonstrated moderate overall agreement. This indicates that utilization of the RSNA classification system agrees with COVID-19 status as determined by the established method of nasal swab PCR. Of note, there was a significant amount of COVID-19 positive patients categorized as “Indeterminate.” These results can be explained by the similarity in the classification criteria between the two categories and presence of mixed imaging features of COVID-19.

The study is limited by its retrospective design and single institutional analysis. Experience among participating radiologist was widely varied and ranged from 3 to 35 years in the participants in this study. This would explain why participants had some difficulty in determining whether to categorize CT findings as “Indeterminate” category or the “Typical” category. The potential solution to this dilemma would be emphasizing the importance of the RSNA classification system prior to our survey other than their own acquired experience with reading chest CTs during the peak of COVID-19 cases. This is a potential limitation of the study as the lack of familiarity with the RSNA classification system may lead to misinterpreting or missing COVID-19 specific imaging findings, this would be expected to affect nonchest-trained participants to a greater degree.

When correlating the interobserver agreement results to the PCR-confirmed COVID-19 status of the patients, this study demonstrated moderate overall agreement. This indicates that utilization of the RSNA classification system agrees with COVID-19 status as determined by the established method of nasal swab PCR. Of note, there was a significant amount of COVID-19 positive patients categorized as “Indeterminate.” These results can be explained by the similarity in the classification criteria between the two categories and presence of mixed imaging features of COVID-19.

Multiple theories explain the overall fair agreement seen in the “Indeterminate” category for the interobserver survey. One major difference between the “Typical” and “Indeterminate” categories is that the former requires rounded and peripherally distributed GGO, while the latter typically demonstrates nonrounded or randomly distributed GGOs. In cases where a mixture of features for different categories was present, participants tended to disagree or split answers between two of the prominent features (Fig 6). It is possible that accurate recognition of which of the two features was present, as well as the presence of both features in a CT, posed difficulty to the
chest fellowship trained group and 2–27 years in the nonchest fellowship-trained group. The variation in years of experience can create mixed results especially in cases with several or little to no findings (Fig 7). There are several articles that discuss the dramatic changes in chest CT findings in COVID-19 patients over time (15–18). Our study did not adequately explore this phenomenon, as the studies included in our survey represent a single moment in the clinical course of the affected patients. Given this constraint, some patients may have presented earlier in the course of their disease without any image findings of COVID-19 and some may have presented later in the course of the disease with the most severe findings. The variation in time from symptom onset to chest CT likely skews the results.

Recently published literature on a new classification system called CO-RADS proposed by Prokop et al (19) utilizes a six-point scale that is based on similar imaging findings discussed among expert RSNA consensus. The CO-RADS system was not introduced to our hospital system during data acquisition, and given the lack of familiarity, was not included within our study.

In conclusion, this study demonstrates moderate overall interobserver agreement and moderate to substantial intraobserver agreement for chest CT findings of COVID-19 based on the RSNA classification system. There is reliable utilization of the RSNA classification system criteria by radiologists when reporting on chest CT findings related to COVID-19.

AUTHOR CONTRIBUTIONS
All authors substantially contributed to the conception or design of the study, the writing and/or revision of the manuscript, approved the final version of the manuscript, and are accountable for the manuscript's contents.

FUNDING
No funding was used for the completion of this study.

DISCLOSURES
The authors have nothing to disclose.

STATEMENT OF DATA INTEGRITY
The authors declare that they had full access to all the data in this study and the authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

ACKNOWLEDGMENTS
Priyanka Annigeri, MD
Vladimir Starcovic, MD
Ting Li, MD
Joshua Joseph, MD
Hannah Saad, MD

Thomas Keimig, MD
David Spizarny, MD
Daniel Myers, MD
Alexander Tassopoulos, MD
Matthew Rheinboldt, MD
John Blase, MD
Milan Pantelic, MD

REFERENCES