European Radiology (2022) 32:3513–3524
https://doi.org/10.1007/s00330-021-08435-2

CHEST

Comparison of semiquantitative chest CT scoring systems to estimate severity in coronavirus disease 2019 (COVID-19) pneumonia

Akitoshi Inoue[1,2] · Hiroaki Takahashi[2] · Tatsuya Ibe[3] · Hisashi Ishii[3] · Yuhei Kurata[3] · Yoshikazu Ishizuka[4] · Yoichiro Hamamoto[3]

Received: 13 April 2021 / Revised: 7 October 2021 / Accepted: 23 October 2021 / Published online: 12 January 2022

Abstract

Objectives To compare the clinical usefulness among three different semiquantitative computed tomography (CT) severity scoring systems for coronavirus disease 2019 (COVID-19) pneumonia.

Methods Two radiologists independently reviewed chest CT images in 108 patients to rate three CT scoring systems (total CT score [TSS], chest CT score [CCTS], and CT severity score [CTSS]). We made a minor modification to CTSS. Quantitative dense area ratio (QDAR: the ratio of lung involvement to lung parenchyma) was calculated using the U-net model. Clinical severity at admission was classified as severe ($n=14$) or mild ($n=94$). Interobserver agreement, interpretation time, and degree of correlation with clinical severity as well as QDAR were evaluated.

Results Interobserver agreement was excellent (intraclass correlation coefficient: 0.952–0.970, $p<0.001$). Mean interpretation time was significantly longer in CTSS (48.9–80.0 s) than in TSS (25.7–41.7 s, $p<0.001$) and CCTS (27.7–39.5 s, $p<0.001$). Area under the curve for differentiating clinical severity at admission was 0.855–0.842 in TSS, 0.853–0.850 in CCTS, and 0.853–0.836 in CTSS. All scoring systems correlated with QDAR in the order of CCTS ($\rho=0.443$–0.448), TSS ($\rho=0.435$–0.437), and CTSS ($\rho=0.415$–0.426).

Conclusions All semiquantitative scoring systems demonstrated substantial diagnostic performance for clinical severity at admission with excellent interobserver agreement. Interpretation time was significantly shorter in TSS and CCTS than in CTSS. The correlation between the scoring system and QDAR was highest in CCTS, followed by TSS and CTSS. CCTS appeared to be the most appropriate CT scoring system for clinical practice.

Key Points

- *Three semiquantitative scoring systems demonstrate substantial accuracy (area under the curve: 0.836–0.855) for diagnosing clinical severity at admission and (area under the curve: 0.786–0.802) for risk of developing critical illness.*
- *Total CT score (TSS) and chest CT score (CCTS) were considered to be more appropriate in terms of clinical usefulness as compared with CT severity score (CTSS), given the shorter interpretation time in TSS and CCTS, and the lowest correlation with quantitative dense area ratio in CTSS.*
- *CCTS is assumed to distinguish subtle from mild lung involvement better than TSS by adopting a 5% threshold in scoring the degree of severity.*

Keywords COVID-19 · SARS-CoV-2 · Severity of illness index · Lung volume measurements · Multidetector computed tomography

✉ Hiroaki Takahashi
h.1982.takahashi@gmail.com

[1] Department of Radiology, Shiga University of Medical Science, Ōtsu, Japan

[2] Department of Radiology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA

[3] Department of Pulmonary Medicine, National Hospital Organization Nishisaitama-Chuo National Hospital, Tokorozawa, Saitama, Japan

[4] Department of Radiology, National Hospital Organization Nishisaitama-Chuo National Hospital, Tokorozawa, Saitama, Japan

Abbreviations

ANOVA	Analysis of variance
AUC	Area under the curve
CCTS	Chest CT score
CI	Confidence interval
COVID-19	Coronavirus disease 2019
CT	Computed tomography
CTSS	CT severity score
ICC	Intraclass coefficient correlation
ROC	Receiver-operating characteristic
RT-PCR	Reverse-transcription polymerase chain reaction
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
TSS	Total CT score

Introduction

Coronavirus disease 2019 (COVID-19), or severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was initially identified in China in 2019 and subsequently spread worldwide, exploding into a pandemic [1]. The respiratory system is the most frequently involved organ in COVID-19 [2]. However, COVID-19 is also known to cause multiorgan system injury, including thromboembolic, neurologic, cardiac, nephrogenic, hepatic, gastrointestinal, endocrinological, and dermatological symptoms, through various pathological mechanisms related to immunity, inflammation, and fibrosis [3, 4]. Computed tomography (CT) plays a crucial role in assessing the severity and extent of lung involvement by COVID-19 [5], whereas the standard confirmatory test for SARS-CoV-2 infection is reverse-transcription polymerase chain reaction (RT-PCR) assay [6]. Characteristic imaging findings of acute lung involvement by COVID-19 include parenchymal or ground-glass opacities with peripherally and lower-lung predominant distribution, crazy-paving appearance, reversed-halo appearance, and subpleural sparing [7–12]. Subpleural sparing along with Hampton's hump and triangular wedge-shaped opacities could suggest the presence of underlying coagulopathies that are commonly seen in COVID-19 patients, and scrutiny on these findings could result in better diagnostic accuracy for COVID-19 infection (Fig. 1) [12]. Interstitial fibrosis is observed as a late sequela of COVID-19 infection [13].

To standardize the subjective assessment of the degree of acute COVID-19 lung involvement, some different semiquantitative CT severity scoring systems have been proposed. The total severity score (TSS), proposed by Li [14], has five grades of severity: 0%, 1–25%, 26–50%, 51–75%, and 75–100% involvement for five lung lobes. The chest CT score (CCTS), proposed by Li [15], is the same as the scoring system developed for severe acute respiratory syndrome [16], which requires diagnostic readers to rate severity by six grades: 0%, < 5%, 5–25%, 26–49%, 50–75%, and > 75% involvement for five lung lobes. The CT severity score (CTSS) proposed by Yang [17] has three severity grades: 0% or absence of involvement, < 50% involvement, and ≥ 50% involvement thresholds, for 20 regions of the lung. Therefore, compared with TSS or CCTS, CTSS requires a simpler grading of severity but, as a trade-off, a more complex assessment for more divided lung regions. Each semiquantitative scoring system should have individual advantages and disadvantages; however, there has been no research to compare the accuracy and efficacy of these different methods.

The purpose of this study is to compare the clinical usefulness of the three different semiquantitative CT severity scoring systems. We evaluate their interobserver agreement, time required for evaluation, and degree of correlation with the clinical severity as well as the computer-calculated quantitative CT severity of the lung involvement.

Materials and methods

Patients

This retrospective study was approved by our institution (National Hospital Organization Nishisaitama-Chuo National Hospital), and written informed consent was waived. We enrolled 108 patients diagnosed with COVID-19 infection by RT-PCR from respiratory tract specimens in our single institution from March 2020 to October 2020. We excluded patients with a history of lung surgery. A pulmonologist (Y.H.) abstracted patient age, sex, body weight, height, body mass index, duration from initial symptom to CT examination, and clinical severity at admission. Clinical severity at admission was classified into binary grades of mild and severe: severe grade was defined as < 93% of percutaneous oxygen saturation or requiring oxygen inhalation. Each patient's risk of developing critical illness was assessed and categorized into three groups (low, moderate, and high) using a predictive scoring system reported by Liang et al. [18].

CT acquisition and reconstruction

Among the 108 enrolled patients, 93 were scanned in our hospital using a 64-row detector CT scanner (Aquilion 64, Canon Medical Systems). We performed scans during inspiratory breath holding using the following parameters: 512–512 matrix, 250–370 mm field of view, and 120 kVp.

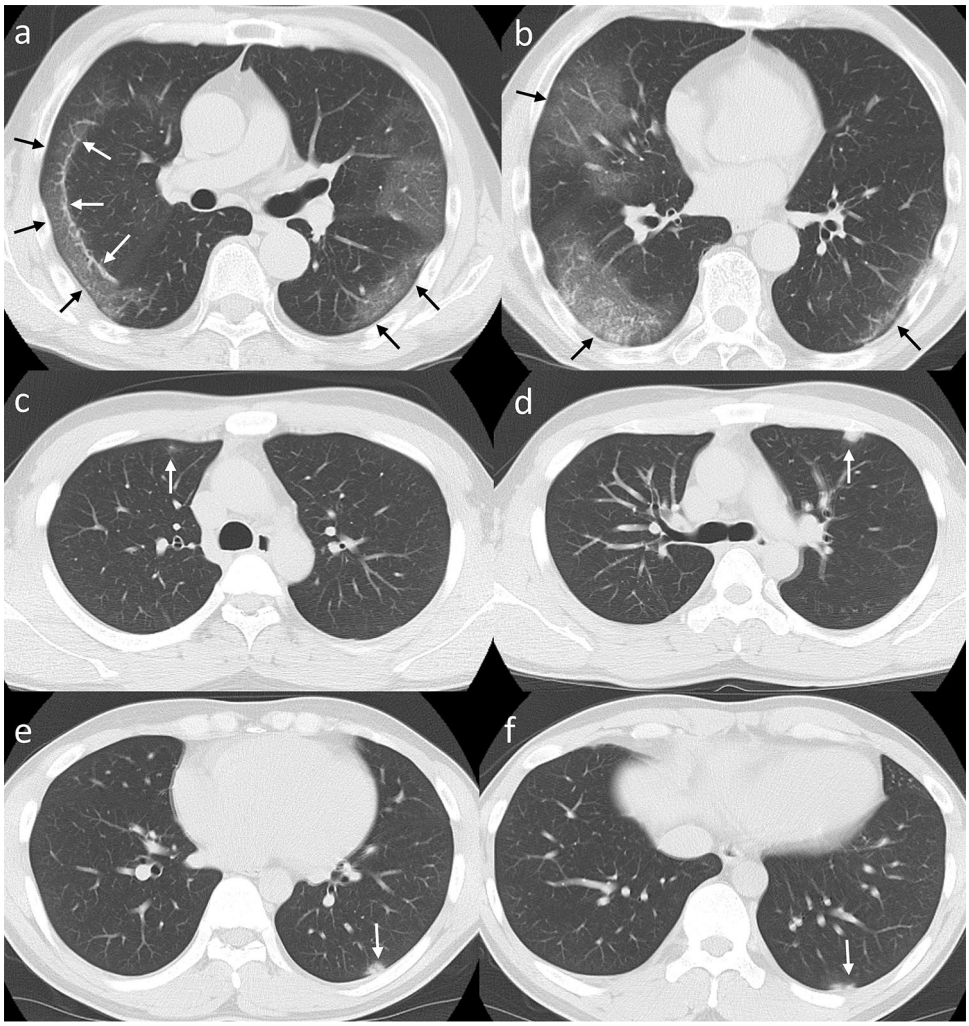

Fig. 1 Chest CT images of COVID-19 in a 65-year-old man ("severe" clinical severity) (**a**, **b**). Subpleural ground-glass opacity is surrounded by dense linear opacity, corresponding with "reversed CT halo sign" (**a** white arrows). The subpleural sparing of ground-glass opacities is noted (subpleural sign: black arrows) [12]. One reader rated a score of 9 in total severity score (TSS), 17 in chest CT score (CCTS), and 22 in CT severity score (CTSS), and the other reader rated a score of 12 in TSS, 16 in CCTS, and 21 in CTSS. Chest CT images of COVID-19 in a 25-year-old man ("mild" clinical severity) (**c**–**f**). The round ground-glass nodule is observed in the peripheral area of the right upper lobe (**a** arrow), and several solid nodules are seen in the peripheral lung of the left upper and lower lobe (**b**–**d** arrows). Both readers rated a score of 1 for the right upper, left upper, and lower lung in TSS and CCTS. In CTSS, one reader rated a score of 1 for the right and left upper lobe and 2 for the left lower lung, whereas the other reader rated a score of 1 for the right upper, left lower, and left lower lung because one of the lesions in the left lower lung was located at the border of segments 6 and 10 (**d** arrow), and one reader rated a score of 1 for two regions

We reconstructed lung setting images with a slice thickness of 5 mm using FC52 kernel. Fifteen patients were scanned outside our hospital using various CT scanners and different parameters. We reconstructed lung setting images with a slice thickness of 5 mm ($n = 14$) or 3 mm ($n = 1$).

Semiquantitative scoring system

Table 1 summarizes three different CT-based semiquantitative scoring systems (TSS, CCTS, and CTSS) assessed in this study. TSS and CCTS were scored using the original methods proposed in previous articles [14, 15]. Scores for TSS and CCTS were rated for five pulmonary lobes. For TSS, scores of 0, 1, 2, 3, and 4 were assigned if parenchymal opacification involved 0%, 1–25%, 26–50%, 51–75%, or 76–100%, respectively. For CCTS, scores of 0, 1, 2, 3, 4, and 5 were rated if parenchymal opacification involved 0%, < 5%, 5–25%, 26–49%, 50–75%, and ≥ 75% (Table 1).

We made a minor modification to CTSS and developed a modified CTSS. The original CTSS has three severity grades: 0% or absence of involvement, < 50% involvement, and ≥ 50% involvement thresholds, for 20 regions of the

Table 1 Summary of the three evaluated semiquantitative scoring systems in this study

	Total severity score (TSS) [13]	Chest CT score (CCTS) [14]	CT severity score (CTSS) [16], modified
No. of evaluated objects	5	5	19
	3 lobes in the right lung	3 lobes in the right lung	10 segments in the right lung
	2 lobes in the left lung	2 lobes in the left lung	9 segments in the left lung
Score for each region	0 (0%), 1 (1–25%), 2 (26–50%), 3 (51–75%), 4 (76–100%)	0 (0%), 1 (<5%), 2 (5–25%), 3 (26–49%), 4 (50–75%), 5 (>75%)	0 (0%), 1 (<50%), 2 (50–100%)
Range of total score	0–20	0–25	0–38
Results in the original study			
Area under curve	0.918	0.870	0.892
Threshold	7.5	7.0	19.5
Sensitivity (%)	82.6	80.0	83.3
Specificity (%)	100	82.8	94.0
Results in this study			
Reader 1			
Area under curve	0.855	0.853	0.853
Threshold	4.5	7.5	12.5
Sensitivity (%)	0.857	0.786	0.857
Specificity (%)	0.809	0.872	0.862
Reader 2			
Area under curve	0.842	0.850	0.836
Threshold	5.5	7.5	13
Sensitivity (%)	0.786	0.786	0.786
Specificity (%)	0.851	0.872	0.851

lung. The original CTSS was intended to make the bilateral lung segments symmetrical and thus subdivided the left apico-posterior segment (S1 + 2) and left anterior basal segment (S8) into two different segments, respectively [17]. However, we encountered cases in which it was difficult to define the border of two regions subdivided from the left S8 as we have scored COVID-19 lung involvement. Therefore, we modified this scoring system using 19 instead of 20 segments, including 10 right-lung segments and 9 left-lung segments, subdividing only the left S1 + 2 into S1 and S2 (Table 1 and Fig. 1).

Reading session

The semiquantitative scores of the three different systems were rated independently by two board-certified radiologists (with 7 and 12 years of experience in thoracic radiology, respectively) in three different sessions. The readers knew only that all patients were positive for COVID-19 infection as confirmed by RT-PCR and were blinded to other clinical information of the patients. The readers rated the score of each patient for each semiquantitative scoring system in three reading sessions, without taking into consideration the confidence level of suspicion for COVID-19 infection. The readers scored the lung lesions only when they thought the findings were related to COVID-19 infection; other lung lesions (e.g., atelectasis and lung nodules/masses) were not considered. Abnormal findings outside the lungs were not described in this session. Readers received prestudy training to rate three sample cases before each session. Interpretation times were recorded in all cases. To avoid recall bias, each reading session was separated by at least 2 weeks.

Automatic quantitative measurement

We performed a quantitative analysis using Python, with a script written by one of the authors (H.T.). The voxel volume of each lung lobe was automatically calculated using U-net (LTRCLobes_R231; model available on GitHub, https://github.com/JoHof/lungmask). The R231 model performs segmentation on individual slices and extracts the right-left lung separately with good performance when dense structures including tumors and consolidation exist. The trachea was not included in the lung segmentation. LTRCLobes performs segmentation of individual lung lobes with limited performance when dense structures exist. The LCRCLober_R231 model runs the R231 and LTRCLobes model and fuses the results [19], in which false negatives from LTRCLobes are filled by R231 predictions and mapped to a neighbor label, whereas false positives from LTRCLobes are removed (Fig. 2a, b).

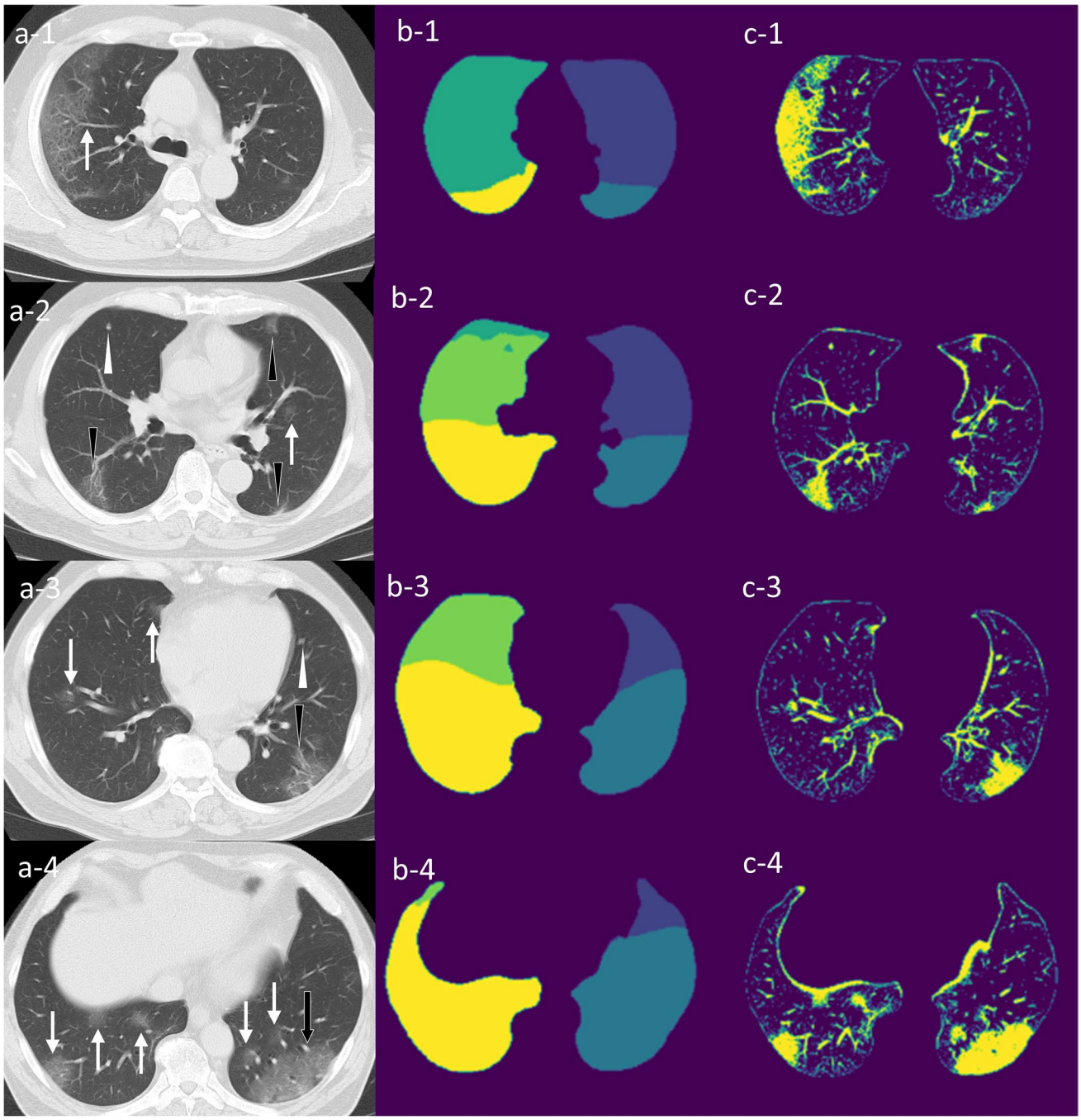

Fig. 2 Masking images of each segmented lobe. Chest CT images of COVID-19 in a 60-year-old man ("severe" clinical severity) show the multiple peripheral rounded opacities abutting the pleura (**a**, 1–4: white arrows). The lesions in the left lower lobe demonstrate bulging opacity presumably indicative of Hampton's hump sign (**a**, 4: black arrow), and triangular wedge-shaped opacities (**a**, 2, 3: black arrowheads). These imaging findings are presumably indicative of infarct [12]. Solid nodules in the right middle lobe and left upper lobe (**a**, 2 and 3: white arrowheads) were considered unrelated to COVID-19. In TSS, both readers recorded 2/1/1/1/2 for the right upper, right middle, right lower, left upper, and left lower lobes, and the patient-level score was 7/20. In CCTS, two readers rated 3/1/2/1/4 and 3/1/2/1/3 for each lung, and the patient-level scores were 11/25 and 10/25, respectively. In CTSS, two readers rated 4/1/5/3/6 and 4/2/5/2/4 for each lobe, and the patient-level scores were 19/38 and 17/38, respectively. The masking images of each segmented lobe (**b**, 1–4) are the corresponding slices of chest CT (**a**, 1–4). The U-net model segments the lung lobes using different colors: right upper lobe, green; right middle lobe, yellow green; right lower lobe, yellow; left upper lobe, dark blue; and left lower lobe, blue green. The images demonstrating the involved areas (**c**, 1–4) are the corresponding slices (**a** and **b**, 1–4). The voxel volumes of areas with CT numbers ranging between − 750 and − 1 were extracted from individual segmented lung lobes (**c**, 1-4). The quantitative dense area ratios were 23.8 in the right upper lobe, 9.0 in the right middle lobe, 21.3 in the right lower lobe, 12.9 in the left upper lobe, and 30.3 in the left lower lobe

The voxel volume of areas with CT values ranging from − 750 to − 1 was extracted from the individual segmented lung lobe(s) (Fig. 2c). The quantitative dense area ratio (QDAR) of the lung lobe(s) was calculated using the following formula:

$$QDAR = \frac{voxel\ volume\ of\ area\ with\ CT\ value\ ranging\ from\ -750\ to\ -1\ in\ lung\ lobe(s)}{voxel\ volume\ of\ lung\ lobe(s)}$$

Statistical analysis

We used intraclass coefficient correlation (ICC) class 2 to assess interobserver agreement of the semiquantitative scoring systems. The agreement outcomes were classified as follows: < 0.50, poor agreement; 0.50–0.75, fair agreement; 0.75–0.90, good agreement, and 0.90–1.00, excellent agreement. To compare the reading time among the three scoring systems, we performed one-way analysis of variance (ANOVA) and paired *t* test with Bonferroni correction to compare differences among groups if the one-way ANOVA revealed a significant difference.

We analyzed the relationship between the three different semiquantitative systems and clinical severity at admission (mild vs. severe) using receiver-operating characteristic (ROC) analysis and compared by DeLong test with Bonferroni correction. The cutoff value was determined with the Youden index. Additionally, the relationship between the three different semiquantitative systems and patients' risk of developing critical illness (low vs. moderate/high) was analyzed using ROC analysis in the same manner.

We analyzed the correlation between three semiquantitative scores and QDAR using Spearman's rank correlation coefficient at both the patient and lobe levels. In the per-patient-level analysis, we evaluated the correlation between the total score of the semiquantitative scale and QDAR. For the lobe-level correlation analysis, the score of the semiquantitative scale was standardized using the following formula:

$$Standerdized\ score\ (TSS,\ CCTS,\ CTSS) = \frac{The\ total\ score\ of\ the\ lobe(s)\ rated\ by\ the\ reader}{The\ expected\ maximum\ total\ score\ of\ the\ lobe(s)}$$

We compared the difference in QDAR among the neighboring lobe-level score categories of 0–4 in TSS and 0–5 in CCTS using a *t* test with Bonferroni correction. A *p* value < 0.05 was considered significant. Statistical analysis was conducted using open-source statistical software (version 3.6.3, R).

Table 2 Patient demographics

Age (years)	46.3 (20.1)
Male	59 (54.6%)
Female	49 (45.3%)
Body weight (kg)	66.0 (16.4)
Height (cm)	164.0 (9.5)
Body mass index (kg/m^2)	24.3 (4.6)
Duration from initial symptom to CT examination (day)	4.8 (3.9)
Clinical severity at admission	
Mild	94 (87.0%)
Severe	14 (13.0%)
Risk of developing critical illness	
Mild	40 (48.8%)
Moderate	41 (50.0%)
Severe	41 (50.0%)

Table 3 Results of semiquantitative scores and automatic quantitative measurement at the patient and lobe levels for each semiquantitative scoring system

	QDAR	Reader 1			Reader 2		
		TSS [standardized]	CCTS [standardized]	CTSS [standardized]	TSS [standardized]	CCTS [standardized]	CTSS [standardized]
Total	21.24 ± 12.53	2.64 ± 3.11 [0.13 ± 0.16]	3.89 ± 4.85 [0.16 ± 0.19]	6.06 ± 7.67 [0.16 ± 0.20]	2.89 ± 3.53 [0.14 ± 0.18]	3.78 ± 4.55 [0.15 ± 0.18]	6.43 ± 8.20 [0.16 ± 0.20]
Right upper lobe	18.20 ± 9.80	0.39 ± 0.67 [0.10 ± 0.17]	0.63 ± 1.03 [0.13 ± 0.21]	0.79 ± 1.24 [0.13 ± 0.20]	0.48 ± 0.71 [0.12 ± 0.18]	0.56 ± 0.91 [0.11 ± 0.18]	0.86 ± 1.34 [0.13 ± 0.21]
Right medial lobe	14.90 ± 8.95	0.35 ± 0.53 [0.09 ± 0.13]	0.45 ± 0.82 [0.09 ± 0.16]	0.56 ± 0.85 [0.16 ± 0.34]	0.40 ± 0.72 [0.10 ± 0.18]	0.49 ± 0.83 [0.10 ± 0.17]	0.56 ± 0.92 [0.17 ± 0.22]
Right lower lobe	24.81 ± 16.91	0.73 ± 0.90 [0.18 ± 0.22]	1.04 ± 1.30 [0.21 ± 0.26]	1.87 ± 2.31 [0.18 ± 0.23]	0.71 ± 0.87 [0.18 ± 0.22]	1.00 ± 1.15 [0.20 ± 0.23]	1.83 ± 2.36 [0.14 ± 0.21]
Left upper lobe	18.99 ± 10.80	0.46 ± 0.66 [0.12 ± 0.16]	0.71 ± 1.00 [0.14 ± 0.20]	1.23 ± 1.66 [0.12 ± 0.19]	0.52 ± 0.76 [0.13 ± 0.19]	0.70 ± 1.00 [0.14 ± 0.20]	1.41 ± 2.18 [0.14 ± 0.22]
Left lower lobe	26.77 ± 18.04	0.69 ± 0.88 [0.17 ± 0.22]	1.06 ± 1.33 [0.21 ± 0.27]	1.67 ± 1.98 [0.20 ± 0.25]	0.78 ± 0.92 [0.19 ± 0.23]	1.02 ± 1.22 [0.20 ± 0.24]	1.77 ± 2.13 [0.21 ± 0.26]

CCTS chest CT score, *CTSS* modified CT severity score, *QDAR* quantitative dense area ratio, *TSS* total CT score

Results

A total of 108 patients (46 ± 20 years old; male:female = 59:49) were enrolled in this study. Body weight, height, and body mass index were measured in 97 patients (89.8%) and were 66.0 ± 16.4 kg, 164.0 ± 9.5 cm, and 24.3 ± 4.6, respectively (Table 2). The duration from initial symptoms to CT examination was available in 104 (96.3%) patients and 4.8 ± 3.9 days. Fourteen patients (13%) had severe clinical severity on admission (i.e., patients who required oxygen inhalation or who had $SpO_2 < 93\%$), and 94 patients (87%) had mild clinical severity (Table 2). Higher scores were observed in the lower lobe than in the upper and middle lobes in all semiquantitative scoring systems by both readers (Table 3).

Table 4 Interobserver agreement of the three semiquantitative scoring systems

	ICC (2, 1)	*p* value
Patient level		
TSS	0.952 (0.928–0.967)	< 0.001*
CCTS	0.970 (0.957–0.979)	< 0.001*
CTSS	0.972 (0.959–0.981)	< 0.001*
Lobe level		
TSS	0.882 (0.861–0.900)	< 0.001*
CCTS	0.936 (0.924–0.945)	< 0.001*
CTSS	0.916 (0.902–0.929)	< 0.001*

CCTS chest CT score, *CTSS* modified CT severity score, *TSS* total CT score, *ICC* intraclass correlation coefficient

*$p < 0.05$

Patient-level interobserver agreement of the three semiquantitative scoring systems showed excellent agreement (ICC: 0.952–0.970, $p < 0.001$). Lobe-level interobserver agreement showed excellent agreement in CCTS and CTSS (0.916–0.936, $p < 0.001$) and good agreement in TSS (0.882, $p < 0.001$; Table 4). The average required time for each case was 25.7 ± 10.2 s for TSS, 27.7 ± 11.7 s for CCTS, and 48.9 ± 28.8 s for CTSS for reader 1 and 41.7 ± 14.9 s for TSS, 39.5 ± 11.7 s for CCTS, and 80.0 ± 37.7 s for CTSS for reader 2. One-way ANOVA indicated a significant difference among the three scoring systems for both readers ($p < 0.001$). In the pairwise comparison using a *t* test, CTSS required significantly more time than TSS and CCTS did in both readers ($p < 0.001$).

Table 1 shows the respective sensitivity, specificity, and cutoff values as calculated by the Youden index for clinical severity at admission. There was no significant difference in AUC for the clinical severity at admission among the three semiquantitative scoring systems for both readers (Table 1 and Fig. 3).

The risk of developing critical illness was assessed in 76% of patients (82/108); the risk could not be calculated in the remaining 24% of patients (26/108) due to the lack of one or more necessary clinical variables. Among the 82 patients, 49% (40/82), 50% (41/82), and 1% (1/82) were categorized as having a low, moderate, or high risk of developing critical illness, respectively. AUC for differentiating the risk of developing critical illness (low vs. moderate/high) of TSS, CCTS, and CTSS were 0.792, 0.818, and 0.786 in reader 1 and 0.788, 0.802, and 0.792 in reader 2, respectively

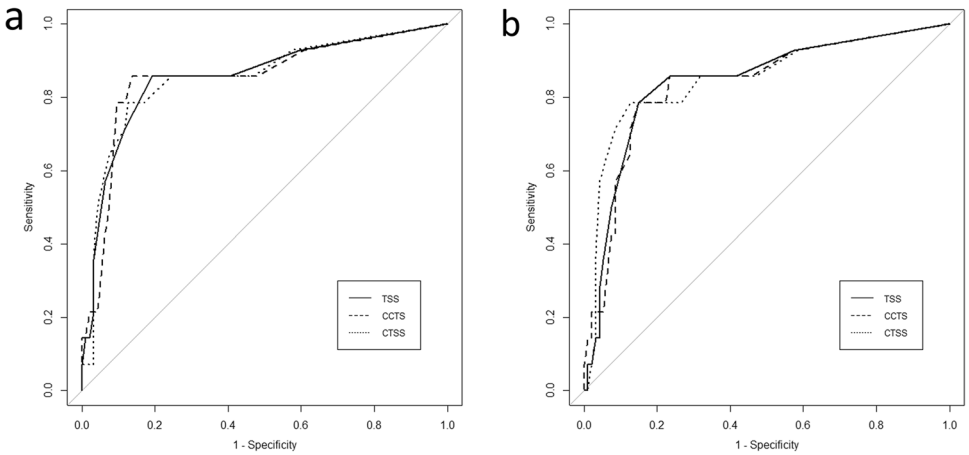

Fig. 3 Receiver-operating characteristic curves for the clinical severity at admission by semiquantitative scoring systems. The receiver-operating characteristic curve is almost similar among the three semiquantitative scoring systems, and the areas under the curve of TSS, CCTS, and CTSS are 0.855 (95% CI 0.732–0.979), 0.853 (95% CI 0.729–0.978), and 0.853 (95% CI 0.726–0.980) for reader 1 (**a**) and 0.842 (95% CI 0.721–0.963), 0.850 (95% CI 0.723–0.977), and 0.836 (95% CI 0.713–0.960) for reader 2 (**b**), respectively. **a** Reader 1, **b** reader 2. CCTS, chest CT score; CTSS, modified CT severity score; TSS, total severity score

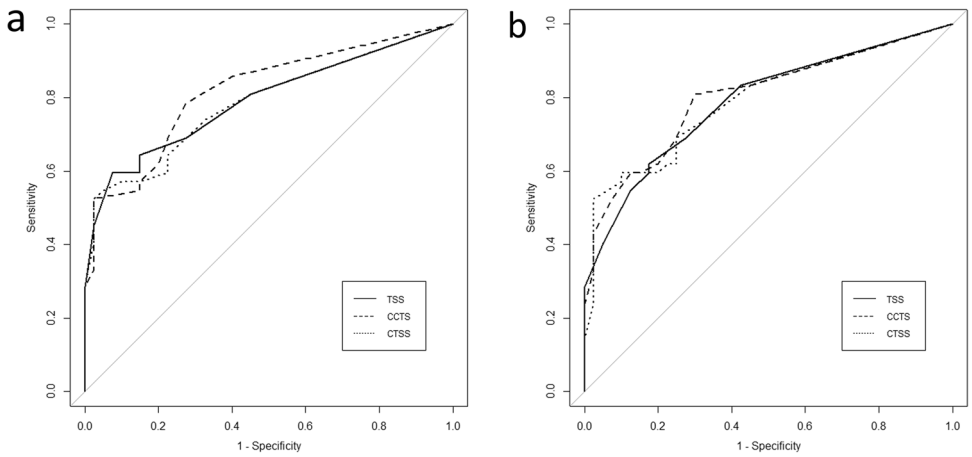

Fig. 4 Receiver-operating characteristic (ROC) curves for predicting the risk of developing critical illness using the semiquantitative scoring systems. The ROC curves were very similar among the three semiquantitative scoring systems, but CCTS demonstrated the highest area under the curve (R1: 0.818 [95% CI: 0.728–0.907] and R2: 0.802 [95% CI: 0.708–0.896]) compared to TSS (R1: 0.792 [95% CI: 0.697–0.888] and R2: 0.788 [95% CI: 0.693–0.883]) and CTSS (R1: 0.786 [95% CI: 0.689–0.883] and R2: 0.792 [95% CI: 0.696–0.888]). **a** Reader 1, **b** reader 2. CCTS, chest CT score; CTSS, modified CT severity score; TSS, total severity score

Table 5 Correlation between the three semiquantitative scoring systems and automatic quantitative measurement

	Reader 1		Reader 2	
	ρ	p value	ρ	p value
Patient level				
TSS	0.437	< 0.001	0.435	< 0.001*
CCTS	0.448	< 0.001	0.443	< 0.001*
CTSS	0.426	< 0.001	0.415	< 0.001*
Lobe level				
TSS	0.392	< 0.001	0.385	< 0.001*
CCTS	0.385	< 0.001	0.415	< 0.001*
CTSS	0.385	< 0.001	0.378	< 0.001*

CCTS chest CT score, *CTSS* modified CT severity score, *TSS* total CT score

*$p < 0.05$

(Fig. 4). There were no significant differences among the three semiquantitative scoring systems for both readers.

All three semiquantitative scoring systems were significantly well correlated with the QDAR for both patient-level correlation and lobe-level correlation (Table 5 and Fig. 5). For the patient level, CCTS showed the highest correlation with the QDAR, followed by TSS with the second highest correlation and CTSS with the lowest correlation for both readers. Five cases showed a QDAR > 50, and three out five cases had a low semiquantitative score (0–1) in each scoring system. These three cases showed mild diffuse mosaic-like increased attenuation in the lung parenchyma, possibly due to air trapping or presumed pulmonary embolus (Westermark sign) [12] or inadequate inspiration.

For lobe-level analysis, the median QDAR was 14.2 and 14.3 in TSS score 0, 16.9 and 16.1 in TSS score 1, 33.4 and 29.5 in TSS score 2, and 53.9 and 53.9 in TSS score 3 by both readers, respectively. A significant difference in QDAR was observed between TSS scores 1 and 2 and between TSS scores 2 and 3 for both readers 1 and 2. The median QDAR was 14.2 and 14.3 in CCTS score 0, 14.2 and 14.0 in CCTS score 1, 22.4 and 21.5 in CCTS score 2, 29.1 and 42.1 in CCTS score 3, and 54.8 and 51.9 in CCTS score 4. We observed a significant difference in QDAR between CCTS scores 1 and 2 and CCTS scores 2 and 3 in both readers 1 and 2 and CCTS scores 3 and 4 in reader 1 (Fig. 6).

Discussion

We compared the clinical usefulness among the three semiquantitative CT-based scoring systems (TSS, CCTS, and CTSS) using the calculated CT severity of the lung (QDAR) as well as clinical severity at admission. Interobserver agreement among the three scoring systems was excellent for the patient level (ICC: 0.952–0.970) and good to excellent for the lobe level (ICC: 0.882–0.936) between the two board-certified radiologists. However, CTSS required a significantly longer time for both readers (R1: 48.9 ± 28.8 s, R2: 80.0 ± 37.7 s) as compared with TSS (R1: 25.7 ± 10.2 s, R2: 41.7 ± 14.9 s, $p < 0.001$) or CCTS (R1: 27.7 ± 11.7 s, R2: 39.5 ± 11.7 s, $p < 0.001$). The AUC in the ROC analysis to predict the clinical severity at admission was 0.842–0.855 in TSS, 0.850–0.853 in CCTS, and 0.836–0.853 in CTSS. The correlation between the scoring system and QDAR

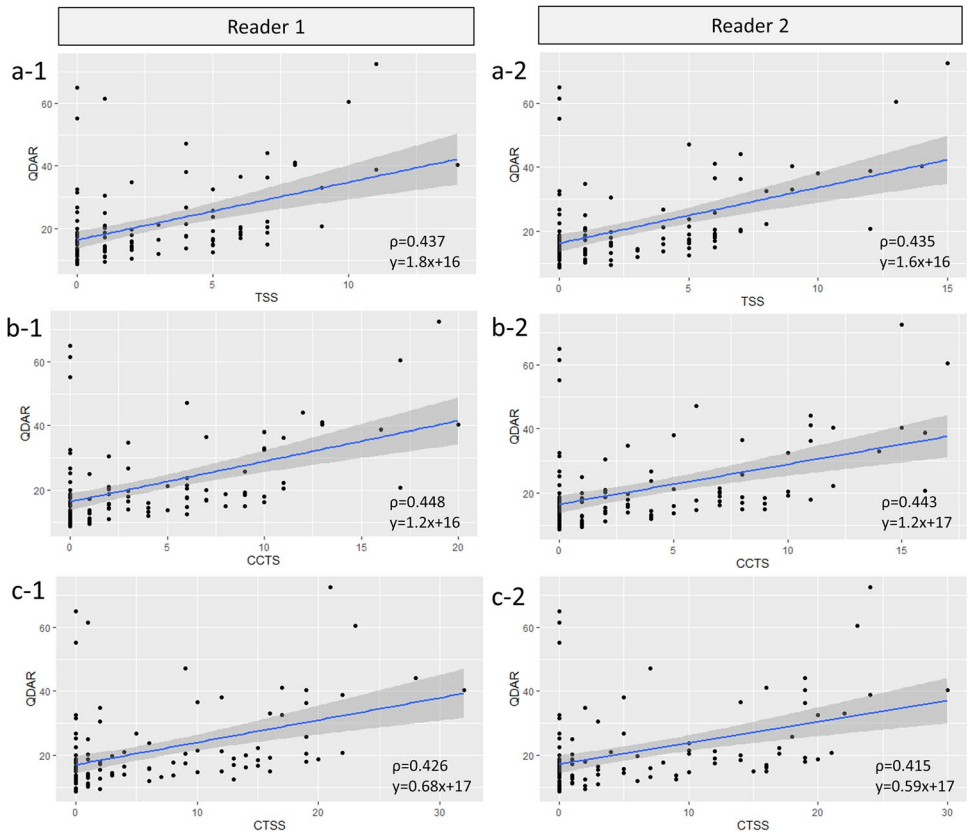

Fig. 5 Patient level of scatterplot and regression line between semiquantitative score and QDAR. The patient level of the scatterplot and regression line between the semiquantitative scores (TSS, CCTS, CTSS) by regression equation and rho value and QDAR for reader 1 (**a**, **b**, **c**—1) and 2 (**a**, **b**, **c**—2). CCTS, chest CT score; CTSS, CT severity score; QDAR, quantitative dense area ratio; TSS, total severity score

was highest in CCTS (0.443–0.448), second highest in TSS (0.435–0.437), and lowest in CTSS (0.415–0.426).

To establish a surrogate standard reference for the CT severity of the lung, we adopted the previously reported U-net model for automated lung lobe segmentation (LTRCLobes_R231). Using this model, we successfully created accurate lung lobe masks bilaterally. We then extracted the additional masks using a CT value ranging from − 750 to − 1 to include both parenchymal and ground-glass opacities, which can commonly be seen in COVID-19 infection, and then calculated the QDAR. The QDAR's major advantage is its ability to provide an accurate and reproducible reference value that could correlate with CT severity rather than human interpretation [20, 21]. Its disadvantage is that it cannot distinguish the qualitative difference within each lobe and therefore should inevitably include false-positive structures within the mask, including pulmonary vasculature, atelectasis, old inflammatory change, fibrotic changes, and inadequate inspiration or air trapping. We consider that some of the higher QDARs seen in the low semiquantitative scaling system should reflect these false positives. In fact, three cases with a QDAR > 50 with low semiquantitative score (0–1) had mild diffuse increased attenuation in the lung parenchyma, possibly because of air trapping or inadequate inspiration. It is also presumed that some of the cases demonstrated Westermark's sign, a sign of pulmonary embolus that appears as heterogeneous attenuation of the lung parenchyma [12].

CTSS was originally developed to investigate the distribution of lung involvement of COVID-19 pneumonia, with both lungs divided equally, resulting in scoring 20 segments in both lungs for 10 segments for each [17]. We modified this scoring system by using 19 instead of 20 segments, including 10 right-lung segments and 9 left-lung segments (details provided in the "Materials and methods" section). CTSS requires readers to evaluate more subdivided regions (20 regions in the original CTSS and 19 regions in CTSS we adapted in this research) with a smaller scale (3 points, 0–2), as compared with TSS (five regions, 5-point scale) and CCTS (five regions, 6-point scale). We assume that the shorter interpretation time in TSS and CCTS as compared with CTSS is mainly accomplished by the smaller interpretation burden in assessing the extent of disease. Given the pandemic situation of COVID-19, physicians need to promptly assess the disease severity of many patients. Furthermore, the AUC of CTSS for clinical severity on admission was similar to that of TSS and CCTS, but the correlation between the scoring system and QDAR was lowest in CTSS. Thus, TSS and CCTS are more appropriate in terms of clinical usefulness as compared with CTSS.

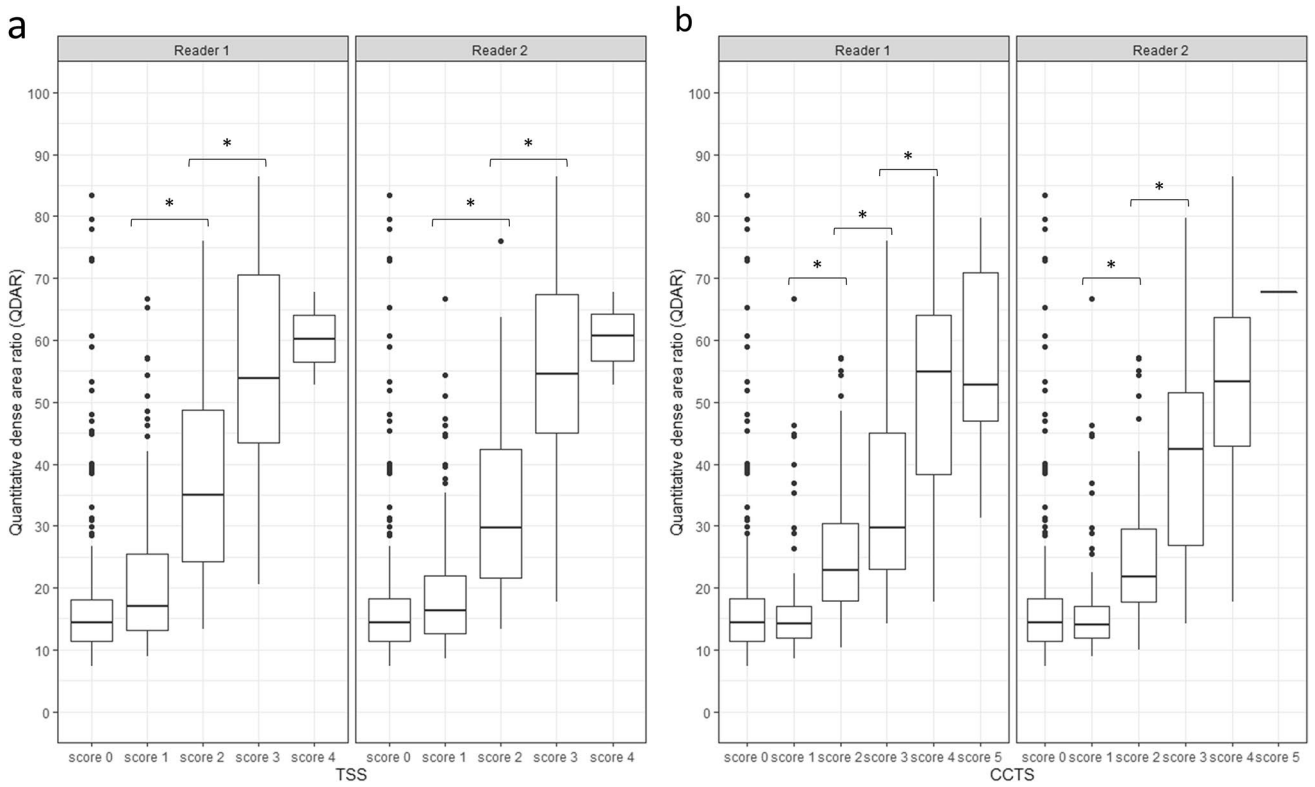

Fig. 6 Lobe-level correlation of the semiquantitative system and QDAR. Both TSS and CCTS demonstrated a proportional correlation to the quantitative dense area ratio in scores of excluding minimal and maximum scores. CCTS, chest CT score; QDAR, quantitative dense area ratio; TSS, total severity score. *$p < 0.05$

The difference between TSS and CCTS relies only on the absence or presence of the 5% threshold in scoring the degree of severity. Therefore, CCTS, which has 5% threshold, is assumed to have a better capability of distinguishing subtle lung involvement from mild lung involvement as compared with TSS which does not have 5% threshold. To quantify this difference, we evaluated lobe-level QDAR in TSS and CCTS and compared neighboring scores (Fig. 6). The median QDAR of score 1 in CCTS (< 5% involvement) was 14.2 in reader 1 and 14.0 reader 2 and that of score 2 in CCTS (5–25%) was 22.4 in reader 1 and 21.5 in reader 2, and the difference in the QDAR between these two scores was significant in both readers. Given that 75% of asymptomatic patients infected with COVID-19 demonstrate small ground-glass opacity in several lobes (1–5 lobes) [10, 22], the category of minimal involvement (< 5%) in CCTS is helpful for stratifying the patients' lung involvement. We presume that the slightly higher correlation with QDAR observed in CCTS compared with TSS should reflect this difference.

Our results are consistent with previous reports demonstrating the three semiquantitative scores predict the clinical severity in COVID-19 pneumonia with substantial sensitivity and specificity [14, 15, 17]. The AUC for clinical severity at admission was almost similar to those of the initial study in TSS (0.842–0.855 vs. 0.819), in CCTS (0.850–0.853 vs. 0.870), and in CTSS (0.853–0.836 vs. 0.892) [14, 15, 17]. The definition of severe clinical severity in this study is almost similar to that of the initial studies, but we did not include partial pressure of arterial blood oxygen or oxygen concentration [14, 15, 17]. The proportion of cases with severe clinical severity (13.0%: 14/108) in our cohort was similar to that in previous studies of TSS (10.3%: 8/78) [14] and CTSS (17.6%: 18/102) [17] but quite different from that in the previous study of CCTS (30.1%: 25/83) [15]. Nevertheless, when validated outside the cohort (this study) with different populations, the diagnostic performance was almost similar. For both readers, there was no significant difference in the AUC for the predictive risk of developing critical illness, but CCTS (0.802–0.818) was higher than TSS (0.792–0.788) and CTSS (0.786–0.792).

This study has some limitations. First, this was a single-center retrospective study. Second, the number of clinically severe patients at admission was small ($n = 12$). However, in terms of the risk of developing critical illness, the ratio of low-risk group patients to moderate/high-risk group patients was approximately 1:1, and CCTS demonstrated the highest AUC for differentiating both the risk of developing critical illness and clinical severity at admission. Third, false-positive structures

were included within the standard reference QDAR, as mentioned above. Fourth, qualitative aspects of pulmonary opacities (i.e., likelihood of COVID-19 infection) were not distinguished in the semiquantitative scoring system. Implementing a model that could automatically score the probability of COVID-19 infection for pulmonary opacities should be investigated in the future. Fifth, CT findings of COVID-19 pneumonia change dramatically over time; ground-glass opacities are dominant immediately after hospitalization [23], whereas consolidation is common within 9–13 days [24]. In our study, the duration between initial symptoms and the CT scan was 4.79 ± 3.91 days, which was the ground-glass opacity dominant phase. Finally, some of the patients underwent CT scan outside our hospital with different CT parameters, including thickness and kernel. This may affect the results of the automated quantification of CT severity.

Conclusion

The three semiquantitative scoring systems (TSS, CCTS, and CTSS) demonstrated substantial diagnostic performances for the clinical severity in patients with COVID-19 pneumonia with excellent interobserver agreement. The interpretation time was significantly shorter in TSS and CCTS than in CTSS. The correlation between scoring system and the QDAR was highest in CCTS, followed by TSS and CTSS. Therefore, we consider CCTS to be the most appropriate CT scoring system for clinical practice.

Acknowledgements The authors would like to thank Enago (www.enago.jp) for the English-language review.

Funding This study has received no finding.

Declarations

Guarantor The scientific guarantor of this publication is Hiroaki Takahashi.

Conflict of Interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and Biometry One of the authors (Hiroaki Takahashi) has significant statistical expertise.

Informed Consent Written informed consent was waived by the Institutional Review Board.

Ethical Approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Performed at one institution

References

1. Chauhan S (2020) Comprehensive review of coronavirus disease 2019 (COVID-19). Biomed J 43:334–340
2. Gavriatopoulou M, Korompoki E, Fotiou D et al (2020) Organ-specific manifestations of COVID-19 infection. Clin Exp Med 20:493–506
3. Lopes-Pacheco M, Silva PL, Cruz FF et al (2021) Pathogenesis of multiple organ injury in COVID-19 and potential therapeutic strategies. Front Physiol 12:593223
4. Gupta A, Madhavan MV, Sehgal K et al (2020) Extrapulmonary manifestations of COVID-19. Nat Med 26:1017–1032
5. Zhou Z, Guo D, Li C et al (2020) Coronavirus disease 2019: initial chest CT findings. Eur Radiol 30:4398–4406
6. Corman VM, Landt O, Kaiser M et al (2020) Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill 25
7. Caruso D, Zerunian M, Polici M et al (2020) Chest CT features of COVID-19 in Rome, Italy. Radiology 296:E79–E85
8. 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
9. Li Y, Xia L (2020) Coronavirus Disease 2019 (COVID-19): Role of chest CT in diagnosis and management. AJR Am J Roentgenol 214:1280–1286
10. Uysal E, Kilincer A, Cebeci H et al (2021) Chest CT findings in RT-PCR positive asymptomatic COVID-19 patients. Clin Imaging 77:37–42
11. Tabatabaei SMH, Talari H, Moghaddas F, Rajebi H (2020) CT Features and short-term prognosis of COVID-19 pneumonia: a single-center study from Kashan. Iran. Radiol Cardiothorac Imaging 2:e200130
12. Merchant SA, Ansari SMS, Merchant N (2020) Additional chest imaging signs that have the potential of being COVID-19 imaging markers. AJR Am J Roentgenol 215:W57–W58
13. Kwee TC, Kwee RM (2020) Chest CT in COVID-19: What the radiologist needs to know. Radiographics 40:1848–1865
14. Li K, Fang Y, Li W et al (2020) CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). Eur Radiol 30:4407–4416
15. Li K, Wu J, Wu F et al (2020) The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Invest Radiol 55:327–331
16. Chang YCYC, Chang SC, Galvin JR et al (2005) Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT. Radiology 236:1067–1075
17. Yang R, Li X, Liu H et al (2020) Chest CT severity score: an imaging tool for assessing severe COVID-19. Radiol Cardiothorac Imaging 2(2):e200047. https://doi.org/10.1148/ryct.2020200047
18. Liang W, Liang H, Ou L et al (2020) Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Intern Med 180:1081–1089
19. Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H, Langs G (2020) Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 4:50
20. Ippolito D, Ragusi M, Gandola D et al (2020) Computed tomography semi-automated lung volume quantification in SARS-CoV-2-related pneumonia. Eur Radiol. https://doi.org/10.1007/s00330-020-07271-0
21. Durhan G, Ardali Duzgun S, Basaran Demirkazik F et al (2020) Visual and software-based quantitative chest CT assessment of

COVID-19: correlation with clinical findings. Diagn Interv Radiol 26:557–564
22. Chang MC, Lee W, Hur J, Park D (2020) Chest computed tomography findings in asymptomatic patients with COVID-19. Respiration 99:748–754
23. Wang Y, Dong C, Hu Y et al (2020) Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: a longitudinal study. Radiology 296:E55–E64
24. Pan F, Ye T, Sun P et al (2020) Time course of lung changes at chest CT during recovery from coronavirus disease 2019 (COVID-19). Radiology 295:715–721