Human-to-AI Interrater Agreement for Lung Ultrasound Scoring in COVID-19 Patients

Noreen Fatima, MSc, Federico Mento, MSc, Alessandro Zanforlin, MD, PhD, Andrea Smargiassi, MD, PhD, Elena Torri, MD, Tiziano Perrone, MD, PhD, Libertario Demi, PhD

Received June 13, 2022, from the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy (NF, FM, LD.); UltraAI, Trento, Italy (NF); Servizio Pneumologico Aziendale, Azienda Sanitaria dell’Alto Adige, Bolzano, Italy (A.Z.); Pulmonary Medicine Unit, Department of Medical and Surgical Sciences, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy (A.S.); Emergency Department, Humanitas Gavazzeni, Bergamo, Italy (E.T., T.P.); and Department of Internal Medicine, IRCCS San Matteo Hospital Foundation, University of Pavia, Pavia, Italy (T.P.). Manuscript accepted for publication June 17, 2022.

LD is a cofounder of UltraAI. NF is a PhD student within a project in collaboration with UltraAI.

Address correspondence to Libertario Demi, PhD, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, Via Sommarive, 9-38123 Povo, Italy.

E-mail: libertario.demi@unitn.it

Objectives—Lung ultrasound (LUS) has sparked significant interest during COVID-19. LUS is based on the detection and analysis of imaging patterns. Vertical artifacts and consolidations are some of the recognized patterns in COVID-19. However, the interrater reliability (IRR) of these findings has not been yet thoroughly investigated. The goal of this study is to assess IRR in LUS COVID-19 data and determine how many LUS videos and operators are required to obtain a reliable result.

Methods—A total of 1035 LUS videos from 59 COVID-19 patients were included. Videos were randomly selected from a dataset of 1807 videos and scored by six human operators (HOs). The videos were also analyzed by artificial intelligence (AI) algorithms. Fleiss’ kappa coefficient results are presented, evaluated at both the video and prognostic levels.

Results—Findings show a stable agreement when evaluating a minimum of 500 videos. The statistical analysis illustrates that, at a video level, a Fleiss’ kappa coefficient of 0.464 (95% confidence interval [CI] = 0.455–0.473) and 0.404 (95% CI = 0.396–0.412) is obtained for pairs of HOs and for AI versus HOs, respectively. At prognostic level, a Fleiss’ kappa coefficient of 0.505 (95% CI = 0.448–0.562) and 0.506 (95% CI = 0.458–0.555) is obtained for pairs of HOs and for AI versus HOs, respectively.

Conclusions—To examine IRR and obtain a reliable evaluation, a minimum of 500 videos are recommended. Moreover, the employed AI algorithms achieve results that are comparable with HOs. This research further provides a methodology that can be useful to benchmark future LUS studies.

Key Words—COVID-19; interobserver agreement; lung ultrasound; SARS-CoV-2

A global pandemic has been triggered by the novel severe acute respiratory syndrome coronavirus (SARS-CoV-2) that causes coronavirus disease 2019 (COVID-19). This acute infectious disease can cause a variety of symptoms, ranging from asymptomatic or moderate flu-like sickness to severe pneumonia, multiorgan failure, and even death. Because of its widespread availability and affordable cost, ultrasound (US) examinations have significant advantages including real-time imaging and use of nonionizing radiations compared to other imaging technologies (X-rays and computed tomography), effectively allowing more patients to benefit from this type of lung imaging. In addition, lung US (LUS) has emerged as a noninvasive method for the rapid assessment of pulmonary illnesses over the past two decades. This technology allows doctors to easily examine patients at the bed side.
even those who are in serious conditions. It is a useful method for detecting and monitoring lung involvement in COVID-19 patients, while also reducing the risk of infection as these portable devices can be easily sanitized after the patient’s examination. The most relevant patterns that can appear in LUS images include horizontal and vertical artifacts (Figure 1, bottom). While horizontal artifacts (A-lines) correlate with a healthy lung surface, vertical artifacts (B-lines) occur when local alterations appear along the lung surface.

Several LUS protocols are based on the detection of these LUS patterns. For example, this study utilized a standardized LUS protocol based on 14 scanning areas and a 4-level scoring system. Specifically, each area is scored from 0 to 3 according to ultrasound aeration patterns (Figure 1, top). Score 0 is associated with a healthy lung surface and consists of a continuous pleural line with horizontal artifacts. Score 1 is assigned when vertical artifacts appear along with an indented pleural line. Score 2 indicates small consolidation areas that appear below a broken pleural line. Score 3 is associated with the presence of white lung with or without consolidations, extending for at least 50% of the pleural line. This acquisition protocol and scoring system have been used in this study to evaluate the interobserver agreement.

LUS is an operator-dependent modality that can provide immediate information on COVID-19 patient conditions. In this prospective observational study, through the collaboration of medical doctors and technical experts, the interobserver agreement for the interpretation of LUS findings in COVID-19 patients was evaluated. The study’s goal is to assess the dependence of the level of agreement on image findings on the number of analyzed LUS videos and operators (2–6). Indeed, this dependence has not been yet thoroughly investigated and previous studies often report results based on a limited amount of data. Moreover, this study is not limited to the assessment of HOs’ interrater agreement, but expands to include the evaluation of interrater agreement between HOs and AI algorithms. These results provide a methodology that can be useful to benchmark future studies on LUS.

Materials and Methods

This study was approved by the Ethical Committee of the Fondazione Policlinico Universitario San Matteo (protocol 20200063198), of the Fondazione Policlinico Universitario Agostino Gemelli, Istituto di Ricovero e Cura a Carattere Scientifico (protocol 20200063198).
0015884120 ID 3117), and of Milano area 1, the Azienda Socio-Sanitaria Territoriale Fatebenefratelli-Sacco (protocol N0031981). The study is part of a registered protocol (NCT04322487). All patients gave informed written consent.

Data Collection
A total of 100 patients (41 female and 59 male), diagnosed as COVID-19 positive through a reverse transcription polymerase chain reaction swab test, were examined by 4 LUS clinical experts (>10 years of experience) using the 14-areas acquisition protocol and 4-level scoring system proposed by Soldati et al. As a subgroup of 33 patients was examined multiple times, a total of 133 LUS examinations (94 performed within the Fondazione Policlinico San Matteo, Pavia, 20 within the Lodi General Hospital, Lodi, and 19 within the Fondazione Policlinico Universitario Agostino Gemelli, Rome) were performed. This resulted in 1807 LUS videos (367,263 frames) acquired using convex probes and different scanners (Esaote MyLab Twice, Esaote MyLab 50, Esaote MyLab Alpha, Philips 1 U22, Esaote MyLab Sigma, and MindRay TE7). The imaging depth was set from 8 to 12 cm and the imaging frequency from 3.5 to 6.6 MHz. From this dataset, we have randomly selected and evaluated the interobserver agreement on a subset of 1035 videos, obtained from 59 COVID-19 patients.

Statistical Analysis
Overall, six LUS experts (H0s) participated in the rating of LUS videos. The first two operators (H0 1 and H0 2) have a technical background with more than 10 and more than 2 years of experience in LUS, respectively, whereas the other four operators (H0 3–6) are clinicians with more than 10 years of experience in LUS. In addition, the agreement between HOs and AI algorithms was assessed. The AI algorithms are able to automatically classify the findings on the videos following the 4-level scoring system. In particular, a first AI algorithm assigns scores to each frame of a video. Therefore, to compare video-level scores assigned by HOs and AI, a second algorithm was used to derive a video-level score from a frame-level score (as assigned by algorithm one). The adopted aggregation technique consists of assigning to a video the highest score assigned at least to 1% of frames composing the video. For more details about the employed AI solution, the reader is referred to the literature. The adopted frame-based scoring algorithms are described in Roy et al., while the aggregation technique solution is detailed in Mento et al.

At video level, each video was independently scored by 6 HOs and the AI solution. At the prognostic level, the cumulative score from each examination (14 scanning areas of individual patients) was computed. The cumulative score (as defined in Perrone et al) ranged from 0 to 42. The patients were stratified into 2 categories based on the prognostic value of cumulative scores. If the cumulative score was greater than 24, the patient has a higher risk of clinical worsening. If the cumulative score is less than or equal to 24, the patient was considered to have a lower risk of clinical worsening. The statistical analysis was carried out at both the video and prognostic levels.

As an example, an agreement between two operators equal to 60% means that these two operators share 60% of data having the exact same score. Specifically, 60% of agreement at video level means that the first operator assigned the same score as the second operator in 60% of the videos. The same procedure is applied at prognostic level, where the assigned labels are binary (category 0 if the cumulative score is less than or equal to 24, category 1 if the cumulative score is greater than 24).

The agreement was first evaluated as a function of the number of LUS videos (Figure 2, top). Specifically, considering N (from 0 to 1000) as the amount of analyzed videos, the agreement for 100 different batches was evaluated, each of which consisting of N videos which were randomly selected from the overall dataset (1035 videos). Following this process, the minimum (min), maximum (max), and standard deviation (SD) of agreement for a given amount of analyzed videos was assessed. The agreement was evaluated also as a function of the number of HOs. Specifically, we indicated as “P-operators agreement” the video level agreement between P HOs, with P ranging from 2 to 6. As an example, a 60% of three-operator agreement at video level means that the first operator assigned the same score as the second and third operator in 60% of the videos. Moreover, we performed the same analyses evaluating the
“one-tolerance” agreement (Figure 2, bottom). One-tolerance agreement means that if the operators assign different scores to a video and their scoring difference is less or equal to one, then the score is considered in agreement. As an example, if operator 1 assigns score 1 to a specific video and operator 2 assigns score 2 to the same video, the score is considered in agreement. In contrast, if operator 1 assigns score 1 to a specific video and operator 2 assigns score 3 to the same video, the score is not considered in agreement, as the difference between the assigned scores is greater than 1 point.

To precisely evaluate the agreement between all the possible combinations of two operators, a box plot analysis (Figure 3) was utilized. A total of 1000 videos were used to perform this analysis. As an example, to draw the box plot of operator 1 (Figure 3A), the agreement between 1000 randomly sampled videos evaluated by operator 1 and operator 2 were computed. Then, another 99 different batches, were used each of which consisting of 1000 videos that were randomly selected from the overall dataset, and compute the agreement between operator 1 and 2. Therefore, 100 (number of batches) values of agreement for the comparison between operator 1 and 2 were obtained. The same process was repeated with operators 3, 4, 5, and 6, thus obtaining 500 (5 pairs of operators × 100 batches) values of agreement, which were used to generate the box plot of operator 1 (Figure 3A, first box plot). This box plot allows evaluation of the agreement between operator 1 with respect to all the other 5 operators. Similarly, the same analysis was performed for the operators from 2 to 6, and for AI (Figure 3B). The same analyses were performed considering the one-tolerance agreement (Figure 3, C and D).

A similar analysis was performed to evaluate the prognostic agreement between all the possible combinations of two operators (Figure 4). In this case, the prognostic agreement by considering all the examinations (14 inspected areas for each examination) was evaluated.

To determine the interobserver agreement, the Fleiss’s kappa ($\kappa$) coefficient was estimated. Specifically,
we computed Fleiss’ kappa at both video (Figure 5A) and prognostic (Figure 5B) level, considering two different groups. The first group includes all the HOs, whereas the second group includes all the HOs plus the AI (see Figure 5). To analyze Fleiss’ kappa results we have interpreted the $f$ value with the following criteria. An $f$ value less than zero indicates poor agreement, whereas, an $f$ value between 0.00 and 0.20, indicates slight agreement, between 0.21 and 0.40 fair agreement, between 0.41 and 0.60 moderate agreement, between 0.61 and 0.80, substantial agreement and a value between 0.81 and 1.00 indicates perfect agreement.\textsuperscript{15}

Results

Agreement as a Function of the Number of Human Operators and Videos

Figure 2 shows the agreement as a function of the HOs. The $x$-axis represents the number of videos,

Figure 3. Agreement between different operators at video level. The top box plots (A, B) show the agreement between pairs of human observers (HOs) (A), and AI versus HOs (B). The bottom box plots (C, D) show the one-tolerance agreement between pairs of HOs (C), and AI versus HOs (D). The $x$-axis represents the operator ID (A, C) or AI (B, D), while the $y$-axis represents the agreement at video level. The red lines of each box represent the median values, and the inferior and superior limits of the box are, respectively, the 25th and 75th percentiles. The maxima and the minima of each box are represented by horizontal black lines.
while the y-axis on left side (blue lines) represents the HOs’ agreement, and the y-axis on right side (red lines) represents the SD value between HOs’ agreement. The first 5 plots (top) illustrate the agreement analysis, whereas the second 5 plots (bottom) of the diagram illustrate the one-tolerance agreement analysis. In the first plot (Figure 2, top row) the agreement between pairs of HOs is presented. When evaluating 1000 videos, the SD reaches 6.3%, whereas the variation between the min and max agreement is

**Figure 4.** The box plots illustrate the agreement between pairs of operators at prognostic level. Box plot (A) indicates the agreement value between pairs of human observers (HOs). Box plot (B) illustrates the agreement between HOs and AI. The x-axis in (A) represents the operator ID, whereas, in (B), the x-axis represents AI. The y-axis represents the agreement at prognostic level. The red lines of each box represent the median values, and the inferior and superior limits of the box are, respectively, the 25th and 75th percentiles. The maxima and the minima of each box are represented by horizontal black lines.

**Figure 5.** Graph (A) illustrates Fleiss’ kappa analysis between human observers (HOs) and AI at video level, while graph (B) illustrates Fleiss’ kappa analysis between HOs and AI at prognostic level. In graph (A) the x-axis indicates whether the Fleiss’ kappa is computed considering all the scores (total agreement) or only specific scores (from 0 to 3). In graph (B) the x-axis indicates that the Fleiss’ kappa is computed in terms of prognostic agreement. The y-axis represents Fleiss’ kappa results. All the Fleiss’ kappa values have been computed for two different groups. The first group (blue square) includes all the HOs, whereas the second group (red circles) includes all the HOs and AI.
about 30%. As expected, by increasing the number of operators that need to be in agreement, both the SD and min-max variation decrease. Specifically, when evaluating 1000 videos, a SD of approximately 4%, 2%, 1.8%, and 0.2% is shown when considering the agreement between 3, 4, 5, and 6 HOs, respectively. The min-max variation is approximately 20%, 10%, 5%, and 1% when considering 3, 4, 5, and 6 HOs, respectively. By evaluating the trend of SD as a function of the number of videos, the agreement reaches stability when a minimum of 500 videos are evaluated. When considering all 5 graphs (Figure 2, top), the difference between the SDs computed on 500 and 1000 videos is always below 1%. These results are consistent with that achieved with one-tolerance agreement (Figure 2, bottom).

**Video-Level Agreement Between Human Versus Human, and AI Versus Human**

Figure 3 shows the agreement between pairs of operators at video level. The top box plots show the agreement between pairs of HOs (Figure 3A), and AI versus HOs (Figure 3B). The agreement is 62% (IQR: 53.2–64.8), 61.9% (IQR: 53.8–61.9), 58% (IQR: 49–60), 62.5% (IQR: 53.8–62.5), 62.2% (IQR: 62–63), and 62.8% (IQR: 62.2–65.6), for operators 1, 2, 3, 4, 5, and 6, respectively. On the other hand, the agreement between AI and the 6 HOs is 45.6% (IQR: 44.6–48.55).

The bottom box plots show the one-tolerance agreement between pairs of HOs (Figure 3C), and AI versus HOs (Figure 3D). The one-tolerance agreement is 90.7% (IQR: 85.6–93.2), 91.5% (IQR: 85.3–94), 92% (IQR: 84–94), 91.9% (IQR: 85.6–94.4), 91.9% (IQR: 91.4–92.7), and 94.3% (IQR: 93.9–94.8), for operators 1, 2, 3, 4, 5, and 6, respectively. The agreement between AI and the six HOs is 82.8% (IQR: 77.3–84).

These results demonstrate how AI has a lower median agreement (Figure 3, B and D) compared to HOs (Figure 3, A and C), but a reduced variability (smaller IQR) compared to operators 1, 2, 3, and 4.

**Prognostic-Level Agreement Between Human Versus Human, and AI Versus Human**

Figure 4 shows the agreement between pairs of operators at the prognostic level. Figure 4A indicates the agreement value between pairs of HOs, whereas Figure 4B illustrates the agreement between HOs and AI. The prognostic agreement is 76.54% (IQR: 66.7–76.54), 77.77% (IQR: 72.83–77.77), 83.95% (IQR: 69.13–86.41), 87.6% (IQR: 72.8–87.65), 83.95% (IQR: 77.7–87.65), and 86% (IQR: 77–87), for operators 1, 2, 3, 4, 5, and 6, respectively. The agreement between AI and the six HOs is 78.3% (IQR: 76.54–80.24).

These results show that AI has a higher median agreement (Figure 4B) compared to operators 1 and 2 (Figure 4A), and a significantly reduced variability (smaller IQR) compared to all HOs.

**Fleiss’ Kappa Analysis**

Figure 5A illustrates Fleiss’ kappa analysis between HOs and AI at video level. The x-axis indicates whether the Fleiss’ kappa is computed considering all the scores (total agreement) or only specific scores (from 0 to 3). All the Fleiss’ kappa values have been computed for two different groups. The first group (blue square) includes all the HOs, whereas the second group (red circles) includes all the HOs and AI. For the first group (all 6 HOs), kappa values are 0.632 (95% CI: 0.616–0.647), 0.236 (95% CI: 0.22–0.252), 0.332 (95% CI: 0.316–0.348), 0.577 (95% CI: 0.562–0.593), and 0.464 (95% CI: 0.455–0.473), for scores 0, 1, 2, 3, and all the scores together (total agreement), respectively. For the second group (all 6 HOs and AI), kappa values are 0.57 (95% CI: 0.557–0.584), 0.195 (95% CI: 0.181–0.208), 0.272 (95% CI: 0.259–0.286), 0.513 (95% CI: 0.499–0.526), and 0.404 (95% CI: 0.396–0.412), for scores 0, 1, 2, 3, and all the scores together (total agreement), respectively. These results show how kappa values slightly decreases when considering the second group (all six HOs and AI). Moreover, it is clear how it is difficult to achieve a high agreement for scores 1 and 2, compared to scores 0 and 3.

Figure 5B illustrates Fleiss’ kappa analysis between HOs and AI at prognostic level. For the first group (all 6 HOs), the kappa value is 0.505 (95% CI: 0.448–0.562), whereas, for the second group (all 6 HOs and AI), the kappa value is 0.506 (95% CI: 0.458–0.555). These results show how AI presents performance similar to HOs in terms of prognostic agreement, achieving a moderate agreement.

**Discussion**

LUS is widely adopted for the clinical evaluation of COVID-19 patients. Patters of interest include
horizontal and vertical artifacts, consolidations, and pleural effusion. Different studies aimed at evaluating the interrater reliability of these patterns, but they generally analyzed a smaller number of videos and/or included a low number of operators. For this reason, in this study the interrater agreement in the analysis of LUS COVID-19 data is assessed. Moreover, the number of videos and operators required to obtain a reliable evaluation is investigated. This study utilized a dataset of 1035 videos acquired from 59 COVID-19 patients. These videos were scored by 6 HOs and by a dedicated AI solution, always according to a 4-level scoring system.

As observed in Figure 2, by evaluating the trend of SD (red lines) as a function of the number of videos, the agreement reaches stability when a minimum of 500 videos are evaluated. Specifically, by considering all 10 graphs (Figure 2), the difference between the SDs computed on 500 and 1000 videos is always below 1%, thus highlighting the stability of agreement.

As shown in Figure 3, when evaluating the video-level agreement, the median agreement between pairs of HOs (from 58% to 62.8%) is higher than the median agreement between HOs and AI (45.6%). However, the IQR of AI (44.6–48.55) is smaller than the IQR of operators 1 (53.2–64.8), 2 (53.8–61.9), 3 (49–60), and 4 (53.85–62.5). This highlights how AI has a lower agreement when compared with HOs, but a low variability (only operators 5 and 6 show a lower variability).

On the other hand, when evaluating the prognostic agreement, AI has a higher median agreement (Figure 4B) compared to operators 1 and 2 (Figure 4A) and maintains a significantly reduced variability (smaller IQR) compared to all HOs. This highlights the possibility to exploit AI for automatic stratification of patients, as the performance of prognostic agreement is comparable with HOs and with significantly reduced variability.

Moreover, by evaluating Fleiss’ kappa results (Figure 5), it is clear how scores 1 and 2 have a lower agreement ($\kappa < 0.35$) compared to scores 0 and 3 ($\kappa > 0.50$). This could be explained by the intrinsic difficulty in classifying in-between scores. On the other hand, the Fleiss’ kappa results on prognostic agreement confirm how the performance of the developed AI ($\kappa = 0.506$) are in line with the performance achieved by HOs ($\kappa = 0.505$), thus highlighting the possibility to exploit AI for automatic patient stratification.

Finally, this study showed how all operators, including HO and AI, have a moderate interobserver agreement in the scoring of LUS data.

Conclusion

In conclusion, to examine the interrater agreement and obtain a reliable evaluation, a minimum of 500 videos is recommended. Moreover, the impact of the number of operators included in a study should be carefully considered when performing this kind of analysis. In conclusion, the analysis of LUS data showed a moderate agreement between operators (HO and AI), both at the video and prognostic level. This research provides a methodological approach that can be useful to benchmark future studies on LUS interobserver agreement.

Acknowledgment

Open Access Funding provided by Universita degli Studi di Trento within the CRUI-CARE Agreement.

References

1. Hsu LY, Chia PY, Lim J. The novel coronavirus (SARS-CoV-2) pandemic. *Ann Acad Med Singap* 2020; 49:105–107.
2. Peng M. Outbreak of COVID-19: an emerging global pandemic threat. *J Biomed* 2020; 129:110499.
3. Touw H, Tuinman P, Gelissen H, Lust E, Elbers P. Lung ultrasound: routine practice for the next generation of internists. *Neth J Med* 2015; 73:100–107.
4. Van Sloun R, Demi L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. *IEEE J Biomed* 2019; 24:957–964.
5. Smargiassi A, Soldati G, Torri E, et al. Lung ultrasound for COVID-19 patchy pneumonia: extended or limited evaluations. *J Ultrasound Med* 2021; 40:521–528.
6. Soldati G, Smargiassi A, Inchingolo R, et al. Is there a role for lung ultrasound during the COVID-19 pandemic? *J Ultrasound Med* 2020; 39:1459–1462.
7. Convissar DL, Gibson LE, Berra L, et al. Narrative review article: application of lung ultrasound during the COVID-19 pandemic: a narrative review. J Anesth 2020; 131:345.

8. Mento F, Demi L. On the influence of imaging parameters on lung ultrasound B-line artifacts, in vitro study. J Acoust Soc Am 2020; 148:975–983.

9. Soldati G, Smargiassi A, Inchingolo R, et al. Proposal for international standardization of the use of lung ultrasound for patients with COVID-19: a simple, quantitative, reproducible method. J Ultrasound Medicine 2020; 39:1413–1419.

10. Mento F, Perrone T, Fiengo A, et al. A multicenter study assessing artificial intelligence capability in scoring lung ultrasound videos of COVID-19 patients. Paper presented at: 2021 IEEE International Ultrasonics Symposium; 2021. https://ieeexplore.ieee.org/document/9593821

11. Roy S, Menapace W, Oei S, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. J IEEE Trans Med Imaging 2020; 39:2676–2687.

12. Mento F, Perrone T, Fiengo A, et al. Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: a multicenter study. J Acoust Soc Am 2021; 149:3626–3634.

13. Perrone T, Soldati G, Padovini L, et al. A new lung ultrasound protocol able to predict worsening in patients affected by severe acute respiratory syndrome coronavirus 2 pneumonia. J Ultrasound Med 2021; 40:1627–1635.

14. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76:378.

15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. J Biometrics 1977; 159–174.

16. Lerchbaumer MH, Lauryn JH, Bachmann U, et al. Point-of-care lung ultrasound in COVID-19 patients: inter- and intra-observer agreement in a prospective observational study. Sci Rep 2021; 11:1–11.

17. Kumar A, Weng Y, Graglia S, et al. Interobserver agreement of lung ultrasound findings of COVID-19. J Ultrasound Med 2021; 40:2369–2376.

18. Nazerian P, Volpicelli G, Vanni S, et al. Accuracy of lung ultrasound for the diagnosis of consolidations when compared to chest computed tomography. Am J Emerg Med 2015; 33:620–625.

19. Millington SJ, Armitfield RT, Guo RJ, et al. Expert agreement in the interpretation of lung ultrasound studies performed on mechanically ventilated patients. J Ultrasound Med 2018; 37:2659–2665.