Article

Inter-rater Variability in the Evaluation of Lung Ultrasound on Videos Acquired from COVID-19 Patients

Joaquin L. Herraiz[1,2], Clara Freijo[1], Jorge Camacho[3], Mario Muñoz[3], Ricardo González[4], Rafael Alonso-Roca[5], Jorge Álvarez-Troncoso[6], Luis Matías Beltrán-Romero[7,8], Máximo Bernabeu-Wittel[7,8], Rafael Blancas[9], Antonio Calvo-Cebrián[10], Ricardo Campo-Linares[11], Jaldún Chehayeb-Morán[12], Jose Chorda-Ribelles[13], Samuel García-Rubio[14], Gonzalo García-de-Casasola[15], Adriana Gil-Rodrigo[16], César Henríquez-Camacho[17], Alba Hernandez-Píriz[18], Carlos Hernandez-Quiles[7], Rafael Llamas-Fuentes[19], Davide Luordo[18], Raquel Marín-Baselga[6], María Cristina Martínez-Díaz[20], María Mateos-González[18], Manuel Mendez-Bailon[21], Francisco Miralles-Aguiar[22], Ramón Nogue[23], Marta Nogué[23], Borja Ortiz de Urbina-Antia[24], Alberto Ángel Oviedo-García[25], José M. Porcel[26], Santiago Rodriguez[7], Diego Aníbal Rodríguez-Serrano[27], Talía Sainz-Costa[28, 29, 30], Ignacio Manuel Sánchez-Barrancos[31], Marta Torres-Arrese[15], Juan Torres-Macho[32], Angela Trueba[33], Tomas Villén-Villegas[34], Juan José Zafra-Sánchez[35] and Yale Tung-Chen[6,36,*]

[1] Nuclear Physics Group, EMFTEL and IPARCOS, Universidad Complutense de Madrid, Madrid, Spain; jlopezhe@ucm.es (JLH); cfreijo@ucm.es (CF)
[2] Health Research Institute (IdISSC). Hospital Clinico San Carlos, Madrid, Spain; jlopezhe@ucm.es
[3] Group of Ultrasound Systems and Technologies, Institute of Physical and Information Technologies (ITEFI), Spanish National Research Council (CSIC), Madrid, Spain. j.camacho@csic.es (JC); mario.munoz.prieto@csic.es (MM)
[4] Dasel SL, Arganda del Rey, Spain; ricardo@daselsistemas.com
[5] Centro de Salud Mar Báltico, Madrid, Spain; alonsorocarafael@gmail.com
[6] Internal Medicine Department. Hosp. Univ. La Paz, Madrid, Spain; jorge.alvarez.troncoso@gmail.com (JAT); raquelzgz14@gmail.com (RMB); yale.tung.chen@gmail.com (YTC)
[7] Internal Medicine Department. Hospital Virgen del Rocío. Sevilla, Spain; luism.beltranromero@gmail.com (LMBR); wittel@cica.es (MBW); quiles_es@yahoo.es (CHQ); santiagorodriguezes@gmail.com (SR)
[8] Department of Medicine. University of Seville, Spain; luism.beltranromero@gmail.com (LMBR); wittel@cica.es (MBW).
[9] Department of Medicine, Universidad Alfonso X. Intensive Care Unit. Hospital Universitario del Tajo. Aranjuez, Spain; rafael.blancas@salud.madrid.org
[10] Centro de Salud de Robledo de Chavela. Robledo de Chavela, Spain; acalceb@gmail.com
[11] Emergency Department. Hospital Santa Bárbara. Puertollano, Spain; rcampol@yahoo.es
[12] Emergency Department. Hospital Universitario Clínico de Valladolid. Valladolid, Spain; jalduncm@gmail.com
[13] Internal Medicine Department. Hospital Universitario General de Valencia. Valencia, Spain; pepechorda@gmail.com
[14] Internal Medicine Department. Hospital Santa Marina. Bilbao, Spain; samuelgarciarubio@gmail.com
[15] Emergency Department. Hospital Fundación de Alcorcón. Madrid, Spain; ggcasasolaster@gmail.com (GGdCS); martatorresarrese@gmail.com (MTA).
[16] Emergency Department. Dr. Balmis General University Hospital, Alicante Institute for Health and Biomedical Research (ISABIAL), Alicante, Spain; adri.gil.rodrigo@gmail.com
[17] Internal Medicine Department. Hospital Universitario Rey Juan Carlos. Móstoles, Spain; doctorcesarhenriquez@gmail.com
[18] Internal Medicine Department. Hospital Infanta Cristina. Parla, Spain; ahpiriz@gmail.com (AHP); davide.tdsco@gmail.com (D.T); ma.mateosglez@gmail.com (MMG)
[19] Emergency Department. Hospital Reina Sofia. Córdoba, Spain; rafael.llamas.fuentes@gmail.com
[20] Intensive Care Medicine Department. Hospital Universitario Príncipe de Asturias. Alcalá de Henares, Spain; cmartinezd@yahoo.es
[21] Internal Medicine Department. Hospital Universitario Clínico San Carlos. Madrid, Spain; manuelmenba@hotmail.com
[22] Anesthesiology Department. Hospital Universitario Puerta del Mar. Cádiz, Spain; curromir@gmail.com
[23] Department of Medicine. Universitat de Lleida. Lleida, Spain; rnogueb@gmail.com (RN); martanogue13@gmail.com (MN)
[24] Pneumology Department. Hospital Universitario de Cruces. Barakaldo, Spain; borjet22@hotmail.com
[25] Emergency Department. Hospital de Valme. Sevilla, Spain; albertoaog1972@hotmail.com
[26] Internal Medicine Department. Hospital Universitario Arnau de Vilanova. Lleida, Spain; jporcelp@yahoo.es
[27] Intensive Care Medicine Department. Hospital Universitario Príncipe de Asturias. Alcalá de Henares, Spain; cancabrilla@hotmail.com

[28] General Pediatrics and Infectious and Tropical Diseases Department. Hospital Universitario La Paz, Madrid, Spain; tsainzcosta@gmail.com
[29] Instituto de Investigación Hospital Universitario La Paz - IdiPAZ, 28046 Madrid, Spain. tsainzcosta@gmail.com
[30] Área de Enfermedades Infecciosas del Centro de Investigación Biomédica en Red del Instituto de Salud Carlos III (CIBERINFEC), Instituto de Salud Carlos III, 28029 Madrid, Spain. tsainzcosta@gmail.com
[31] Centro de Salud Pío XII (Ciudad Real I). Ciudad Real, Spain; ignaciomsb@telefonica.net
[32] Internal Medicine Department. Hospital Universitario Infanta Leonor-Virgen de La Torre, Madrid, Spain. E-mail: juan.torresm@salud.madrid.org
[33] Internal Medicine Department. Hospital de Emergencias Enfermera Isabel Zendal. Madrid, Spain. angelatrueba@gmail.com
[34] Department of Medicine. Universidad Francisco de Vitoria. Madrid, Spain. tomasvillen@gmail.com
[35] Emergency department. Hospital San Eloy. Barakaldo, Spain; juanjose.zafrasanchez@gmail.com
[36] Department of Medicine, Universidad Alfonso X. Madrid, Spain; yale.tung.chen@gmail.com

* Correspondence: yale.tung.chen@gmail.com; Tel.: ++34 917 27 70 00

Abstract: Lung ultrasound (LUS) allows the detection of a series of manifestations of COVID-19 such as B lines and consolidations. The objective of this work was to study the inter-rater reliability (IRR) when detecting signs associated with COVID-19 in the LUS, as well as the impact of performing the test in the longitudinal or transverse orientation. 33 physicians with advanced experience in LUS, independently evaluated ultrasound videos previously acquired with the ULTRACOV system of 20 patients with confirmed COVID-19. In each patient, 24 videos of 3 seconds were acquired (using 12 positions with the probe in longitudinal and transverse orientations). Physicians had no information about the patients or other previous evaluations. The score assigned to each acquisition followed the convention applied in previous studies. A substantial IRR was found in the cases of normal LUS (κ = 0.74), only a fair IRR for the presence of individual B lines (κ = 0.36) and for confluent B lines occupying <50% (κ = 0.26), and a moderate IRR in consolidations and B-lines >50% (κ = 0.50). No statistically significant differences between the longitudinal and transverse scans were found. The IRR in LUS of COVID-19 patients may benefit from more standardization of the clinical protocols.

Keywords: Coronavirus disease 2019; interobserver agreement; interrater reliability; lung ultrasound; point-of-care ultrasound; reliability; severe acute respiratory syndrome; ultrasound

1. Introduction

Lung ultrasound (LUS) is used to differentiate quickly and precisely between the most common causes of respiratory problems. LUS has been extensively studied as a bedside diagnostic tool, and it is now universally included in point-of-care ultrasound (PoCUS) guidelines with high-quality supporting evidence [1]. LUS has the potential to refashion healthcare delivery and enables an augmented clinical interpretation of patient's status in real-time, which can have an immediate impact on clinical decisions, and even be used to monitor response to therapy and evolution [2–4]. Also, LUS devices are typically less expensive than conventional radiological equipment such as X-ray or computed tomography (CT) machines, making them ideal for locations with limited access to these resources [5,6].

LUS has demonstrated the ability to provide immediate information on the condition of COVID-19 patients [4,7]. There are multiple pulmonary manifestations of COVID-19 that can be observed with LUS, such as the presence of pleural effusion, B-line artifacts, or consolidations [8–10].

As LUS allows doctors to examine COVID-19 patients at the bedside, even those who are in serious conditions, it has been a convenient method for detecting and monitoring lung involvement, as well as predicting admission to the intensive care unit (ICU) and mortality [3,4,7,11,12]. Furthermore, LUS reduces the risk of infection compared to other imaging modalities as these portable devices can be easily sanitized after the patient's examination [6,13].

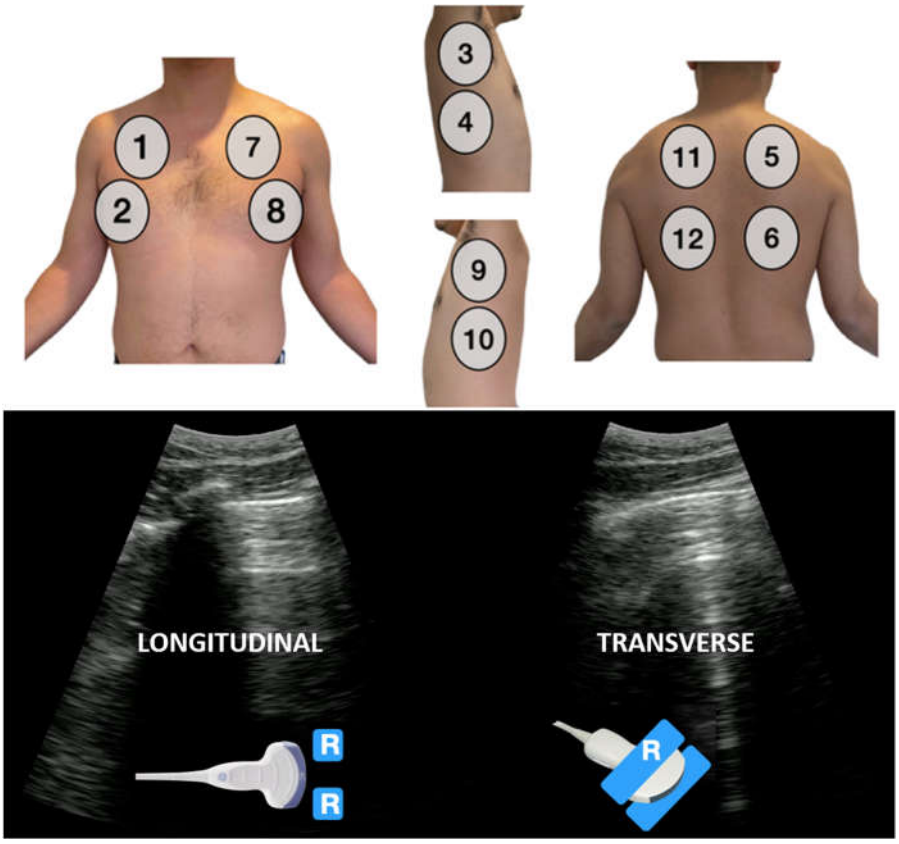


Figure 1. Ultrasound scanning locations (top row) and the two orientations of the transducer considered (bottom row). "R" stands for rib.

However, LUS is an operator-dependent modality, and its utility depends on accurate acquisition and interpretation by bedside physicians [14–17]. Poor image acquisition and incorrect identification and interpretation of artifacts are potential sources of error in the clinical application of LUS [1]. In previous studies, most of them accomplished outside of COVID-19, LUS findings demonstrated moderate to fair interrater agreement. However, as the observed agreement in the interpretation of frequently occurring events may be to some extent due to chance, more studies with a controlled environment are required to determine the accuracy with which physicians can interpret LUS acquisitions.

Furthermore, to the best of our knowledge, there are no published studies to date that specifically evaluate the best orientation of the transducer (i.e., transverse, or longitudinal, see Figure 1) in an LUS acquisition in COVID-19 patients.

In this study, we first characterized the inter-rater agreement of LUS experts when evaluating the main POC-LUS findings for COVID-19. Our hypothesis was that kappa agreement in ultrasound artifact and diagnostic interpretation would be substantial, based on the high agreement in other clinical scenarios. We also evaluated the impact of the transducer orientation in LUS acquisitions in COVID-19 patients on the observed findings.

2. Materials and Methods

In this study, a total of 33 physicians (internal medicine n = 16, intensivist n = 4; family physician n = 5, pneumology = 1, pediatrics = 1, and emergency medicine, n = 6) from 29 different healthcare centers in Spain, with advanced experience in performing and interpreting LUS, evaluated independently previously acquired ultrasound videos of 20 patients. All of them had more than 3 years of performing and interpreting LUS.

The acquisitions corresponded to patients with COVID-19 diagnosed by nasopharyngeal RT-PCR for SARS-CoV-2 obtained in the internal medicine service of two different hospitals in Madrid collected during the summer of 2021 [8]. All scans were collected by 2 of the study authors (YT-C, AT-V). This study utilized a standardized LUS protocol based on 12 scanning areas and a 4-level scoring system [18]. (Figure 1). Specifically, each area is scored from 0 to 3 according to the observed patterns (Figure 2). Score 0 is associated with a healthy lung surface and consists of a continuous pleural line with horizontal artifacts (A-lines). Score 1 is assigned when individual vertical artifacts appear (B-lines) along with an irregular pleural line. Score 2 indicates confluent B-lines in less than 50% of the pleural line. Score 3 is associated with confluent B-lines extending more than 50% of the pleural line, as well as subpleural or lobar consolidation or pleural effusion. In each zone, the ultrasound probe was used in longitudinal and transverse positions, and a 3-seconds video of 20 fps was recorded. In total, 24 videos of 3 seconds each were acquired per patient. No patient had more than one scan in the database.

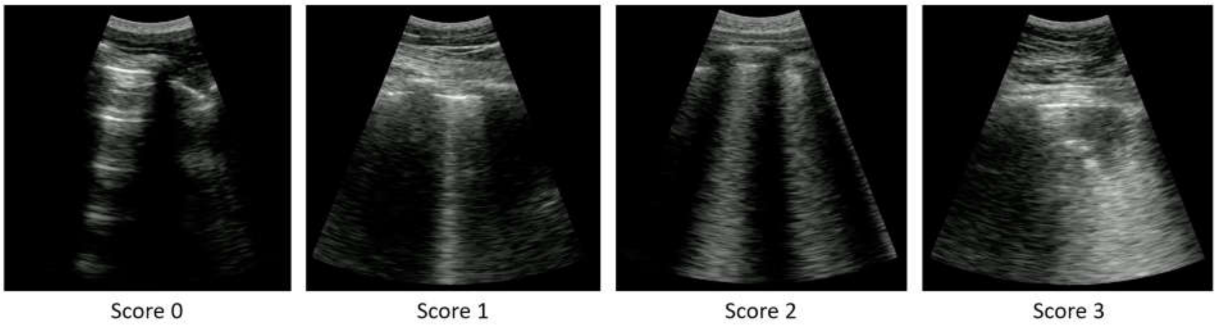


Figure 2. Examples of lung ultrasound images illustrating the four different score levels.

In all cases, the data was acquired with the ULTRACOV ultrasound scanner prototype using a 3.5 MHz convex probe with 128-channel ultrasound electronics [8]. This resulted in a total of 480 videos (28,800 frames). The imaging depth in all cases was set to 13 cm. More details of the data acquisition can be found in [8].

The data was acquired in a study investigating the reduction in the exploration time per patient when using an ultrasound system developed specifically for LUS, which was conducted at a tertiary academic hospital and an emergency field hospital. This study received institutional review board (IRB) approval at both participating sites and all patients' consent was collected.

Several LUS protocols have been proposed for the lung assessment of COVID-19 patients based on the number of areas or points to explore. We adopted a 12-zone scanning protocol which has been previously validated and shown to be consistent with higher ICC and a higher degree of concordance with CT.

The selected patients for this study (n=20) were selected from the total acquired dataset from that study (n=28) so that half of the cases (n=10) corresponded to patients with relatively good condition (with a total score between 1 and 7 based on the in-situ assessment of the LUS expert) and the other half (n=10) corresponded to patients with moderate condition (with a score between 8 and 18 based on the in-situ assessment of the LUS expert). As the LUS device was not located within the ICU, no severe cases were present in the database.

The physicians had no information about the patients in the survey and were blinded to their history and clinical information. They also did not have access to the characteristics of the scanner and the evaluations performed by the other physicians.

The sonographer expert who collected the videos also performed the survey, so that a comparison between the findings obtained during the examination and the ones observed in the surveyed videos was also done. As the survey was performed 9 months after the scans there was no recollection of each patient's status at that time.

2.1. Preparation

The physicians were instructed to independently assess the de-identified studies and provide their interpretation using a survey web. All physicians received instructions at the beginning that contained the scanning protocol and definition of the orientation of the probe in each case. No other instruction on image interpretation, or the definitions for each of the pathological findings was provided during the independent assessment period. The physicians who interpreted the images were blinded to any patient information or previous interpretation. Furthermore, they had no prior experience with the system used to collect the videos.

In previous studies [4,14], physicians met before performing the evaluations for the inter-rater study to review a few sample videos to discuss their real-time interpretations. In this case, no previous calibration session was conducted.

2.2. Data Analysis

We performed a series of statistical analyses comparing the interpretation of the presence of ultrasound artifacts, and the ultrasound diagnosis performed by the physicians. They were all performed with python using NumPy and Scikit-learn libraries.

First, we evaluated the agreement between raters of the individual scores (0,1,2, or 3) assigned by each observer to each of the 480 videos of the study. These videos correspond to the 20 patients, with 12 zones and 2 probe orientations each. Cohen's kappa was used to quantify the interrater agreement between each pair of physicians [19]. Similar to correlation coefficients, it can range from −1 to +1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters. While kappa values below 0 are possible, they are unlikely in practice.

Based on the total score from the evaluation of the 12 zones, patients were classified into 4 subgroups: A. total score 0; B. total score between 1 and 7; C. total score between 8 and 18; D. total score between 19 and 36. These subgroups have been used in previous studies to obtain a fair indication of the severity of their condition. Similarly, to the previous case, Cohen's kappa between this 4-class classification was obtained. In this case, the analysis was performed separately for the longitudinal and transverse examinations. The results are shown in Figure 4.

The agreement in the interpretation of each ultrasound artifact (A-lines, isolated B-lines, confluent B-lines, and consolidations) was also assessed separately. The degree of inter-rater agreement was evaluated using Fleiss' kappa statistics (k) and corresponding 95% confidence intervals. The analysis with the Fleiss kappa was originally described by Cohen, Landis & Koch [19,20] and it is an adaptation of Cohen's kappa for 3 or more raters. k values close to 1 imply strong agreement beyond chance in the LUS diagnosis [19,20]. We interpreted the scaled kappa statistics as follows: k ≤ 0, less than chance agreement; k 0.01–0.20, slight agreement; k 0.21–0.40, fair agreement; k 0.41–0.60, moderate agreement; k 0.61–0.80, substantial agreement; and k ≥ 0.81, near perfect agreement [19,20]. Table 1 contains the results of this analysis.

As an alternative way to visualize the agreement in the findings, Figure 5 shows a matrix of the scores assigned to each video with respect to the most voted score (among the 33 evaluations), which can be considered as a surrogate of the ground truth. This provides a quick view of what are the most challenging scores. Agreement in the comparison of the ultrasound diagnosis performed in situ and with the recorded videos was also done utilizing k adjusted for maximum attainable agreement.

3. Results

3.1. Patients

The selected patients in this study (n=20) corresponded to admissions to the hospital due to COVID-19. The mean age was 53.2 years (standard deviation—SD 11.9) and 45% were female. Five patients (25%) had hypertension, 2 patients (10%) had diabetes, and 0 patients had cardiovascular disease. They had an average of 19.1 days (SD 20.6) after symptom onset, consisting of fever (95%), shortness of breath (75%), and weakness (85%). The mean lymphocyte count was 1.81 × 109 (SD 1.00), C-reactive protein was 29.4mg/dL (SD 33.3), and D-dimer 536.47 ng/mL (SD 315.7) at admission. No patients died at follow-up. The LUS exams were performed within 2-3 days of their hospital admission after obtaining consent. None of these patients ended up in the ICU and were discharged after several days/weeks at the hospital.

3.2. Overall agreement between raters

The overall Cohen's kappa statistics between each pair of raters of the 480 videos are shown in Figure 3. Raters are sorted from left to right based on their overall Cohen's kappa with their peers (which varies from 0.45 to 0.78).

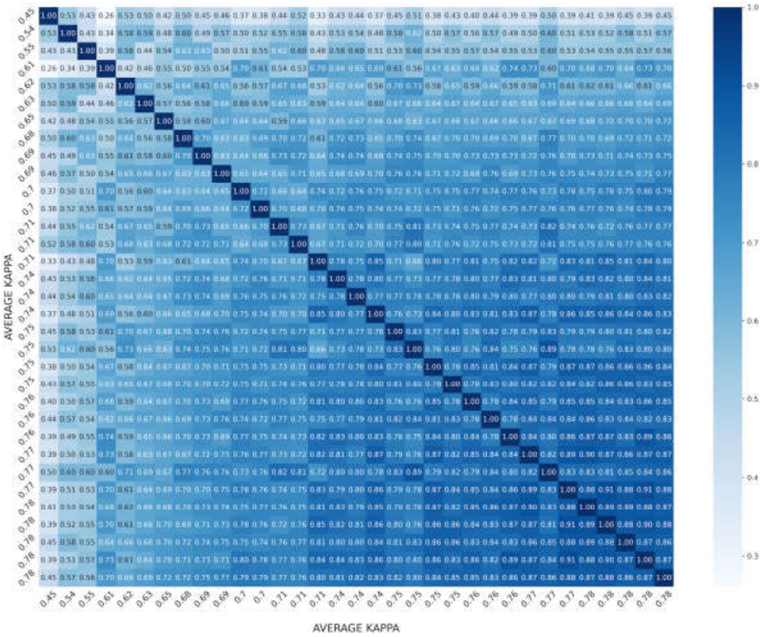
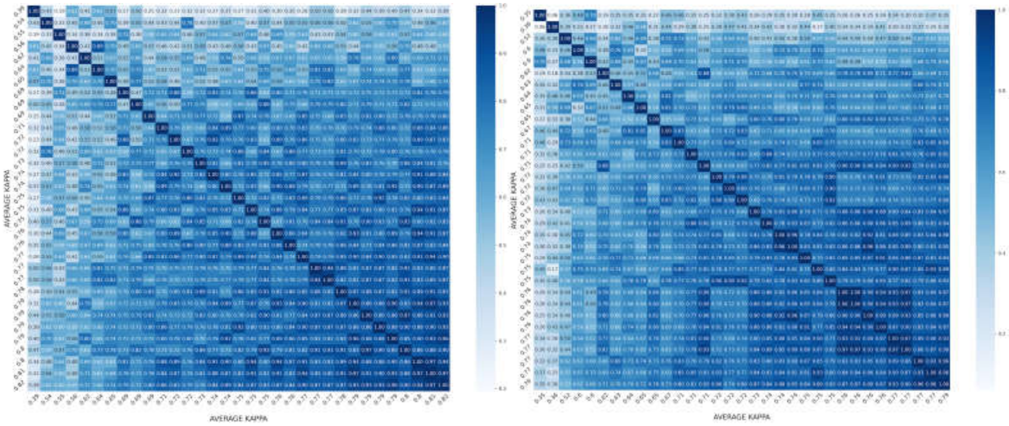


Figure 3. Comparison of the agreement between each pair of raters using Cohen's kappa. It was obtained from the scores assigned by each rater using all 480 videos. Raters are sorted from left to right based on their overall Cohen's kappa with their peers (indicated in the axis ticks).

The comparison of the evaluations of the sonographer who collected the videos performed during the examination with respect to the ones observed in the surveyed videos indicates a Cohen's kappa value of 0.68 (moderate agreement).

The most relevant outcome of the patient LUS evaluation is their classification into 4 subgroups based on the severity of their lung condition. Therefore, we evaluated Cohen's kappa between the classification of patients in each subgroup performed by each physician considering longitudinal and transversal directions (Fig. 4). The agreement was slightly higher with the studies performed in the longitudinal direction (Fig. 4a).



(**a**) (**b**)

Figure 4. Cohen's kappa between raters for (**a**) longitudinal acquisitions; and (**b**) transversal acquisitions classifying patients into 4 subgroups according to their total score. Raters are sorted from left to right based on their overall Cohen's kappa with their peers (indicated in the axis ticks).

3.2. Agreement in specific findings

Regarding the degree of agreement between physicians with respect to the specific finding, Table 1 summarizes Fleiss' kappa analysis. There was a substantial agreement on determining whether a scan contained no abnormalities, in most cases with A-lines (κ = 0.74); A fair interrater agreement on the presence of individual B-lines (κ = 0.36), as well as on the presence of confluent B-lines occupying less than 50% of the ultrasound image (κ = 0.26). And a moderate agreement was found for confluent B-lines occupying more than 50% and consolidations (κ = 0.50).

Table 1. Fleiss Kappa analysis of the interrater agreement in the findings in all the videos.

Score	Finding	**Fleiss Kappa (*k* and 95% CI)**		Agreement
0	Normal / A-lines	0.74	[0.71 - 0.76]	Substantial
1	Individual B-lines	0.36	[0.33 - 0.39]	Fair
2	Confluent B-lines < 50%	0.26	[0.24 - 0.29]	Fair
3	Confluent B-lines > 50% & Consolidations	0.50	[0.47 - 0.53]	Moderate

Figure 5 shows the matrix with the information on how the scores are assigned to each video with respect to the most voted score (mode) in each case. The most voted score (mode) may be considered a good estimation of the ground truth. The largest differences are found for score=2.

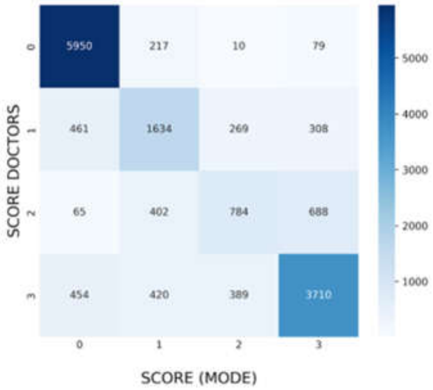
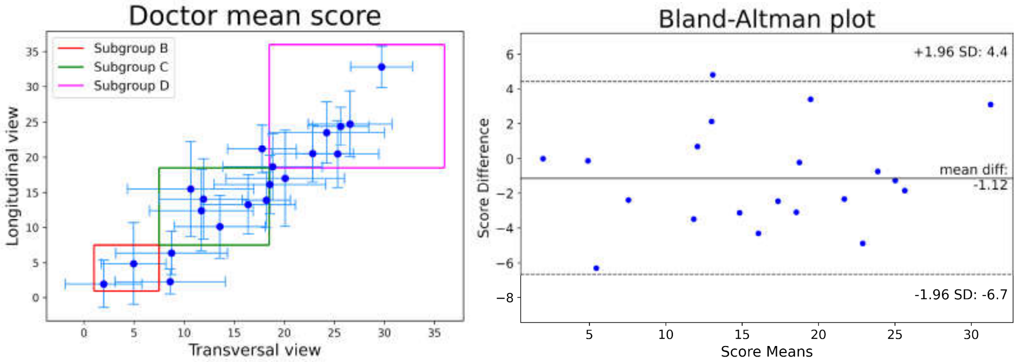


Figure 5. Scores assigned to each video (vertical axis) with respect to the most voted scores (mode, horizontal axis) among the 33 evaluations in each case.

3.2. Agreement in specific findings

Regarding the impact of the probe orientation (longitudinal or transversal, as shown in Fig. 1) when performing the study, the total score assigned to each patient in both cases is shown in Figure 6. A scatter plot shows a very good correlation between both types of examination (R^2=0.87) and the Bland-Altman plot of longitudinal minus transversal scores indicates that on average the longitudinal view provides slightly lower scores (-1.12).



(**a**) (**b**)

Figure 6. (a) Scatter plot with the Longitudinal vs Transversal total scores per patient. Points and error bars correspond to the average and standard deviation respectively of the evaluations obtained from all the physicians. The 3 subgroups shown correspond to the classification of the patients based on their severity; **(b)** Bland-Altman plot of the total score total per patient (obtained as the average value of all the evaluations) assigned to the videos acquired with Longitudinal and Transversal probe orientation. "Score difference" indicates Longitudinal minus Transversal scores.

4. Discussion

Easy-to-access and reliable diagnostic methods which can accurately guide management in COVID-19 are vital in nonhospital settings and areas with limited resources [5]. Some studies start to point out that LUS could be a first-line diagnostic tool alternative to conventional chest X-ray and CT scans since there is no exposure to ionizing radiation [4,6,13]. Moreover, it can be considered in vulnerable populations such as pregnant women and children.

Previous research has shown that COVID-19 has notable LUS characteristics, such as B-lines or consolidations [11]. These findings correlate well with COVID-19 CT findings, such as peripheral nodularity and ground-glass consolidations [9]. As a result, given that LUS may be able to predict outcomes in COVID-19 patients, it is crucial to ascertain whether clinicians can correctly interpret these results.

In this study, several LUS findings demonstrated moderate agreement (e.g., consolidations), and others a fair agreement (individual B-lines and confluent B-lines <50%). Therefore, LUS may represent a reliable diagnostic and prognostic clinical entity for COVID-19. And there was substantial agreement on whether an LUS scan was interpreted as abnormal versus normal. In addition, beyond COVID-19, an abnormal LUS scan has prognostic implications for multiple diseases. This study represents the first investigation of the interobserver agreement of LUS findings in COVID-19 obtained with the same device and including practitioners from multiple specialties and centers, who commonly use different portable devices.

Our results are similar to other previous studies on interrater reliability for LUS outside of COVID-19. Previous investigations have demonstrated moderate to substantial agreement for B-lines [15–17]. In contrast, there is only moderate to a fair agreement for consolidations. This is similar to the results obtained in a previous LUS study with COVID-19 patients [14].

This work shows the importance of working towards a more standardized interpretation protocol. Among the possible solutions, may be considered the following options:

doi:10.20944/preprints202211.0285.v1

1) Standardization of the terminology to describe artifacts and signs in LUS is essential. There are various definitions in the literature, especially for consolidations, but also regarding pleural abnormalities [1], which was not considered in this study. Although consolidations had a moderate agreement in this study, the reliability of this finding might improve with more specific definitions and consensus-based guidelines.

2) The use of automatic tools to quickly analyze the acquisitions and obtain some quantitative values such as the percentage of affected pleura (B lines <50% or >50%) and the size of the consolidations may be helpful to obtain more consistent results among raters.

3) The length of the acquired videos (3 seconds in this study) may be extended to provide more information in some cases.

4) Having additional clinical data about the patients may also help in their evaluation.

There are several limitations to this study. Due to its dynamic nature, the use of LUS is fundamentally different from traditional medical imaging practices in which an exam is performed by a technologist and interpreted remotely by a physician with limited clinical knowledge of the patient. The same provider performs and interprets the study, integrates the findings into the clinical setting immediately, and repeats the study as needed to identify changes associated with bedside interventions. In this case, the raters did not have the opportunity to explore the patients or adapt the ultrasound exploration according to their preferences and findings. Therefore, despite allowing us to evaluate the scans in a very controlled setting (same device, same image quality), this kind of patient observation is not realistic. This fact may have caused some errors in the interpretation of some particular cases. The impact of this was evaluated by performing a comparison of the evaluations of the sonographer who collected the videos during the examination with respect to the ones observed in the surveyed videos. The moderate agreement found (Cohen's kappa 0.68) in this case, is a good indication of the differences that may be expected between in-situ and evaluations performed with a recorded video.

Furthermore, in this study, there were no patients in very severe conditions. This reduced the number and size of the consolidations, making them harder to identify. As shown in Figure 5, and Table 1, the cases with score=2 were the ones with significantly higher disagreements.

Regarding which is the best way to perform LUS, i.e. using longitudinal or transversal view, our results show that there is a very good correlation between both types of examination (R^2=0.87) although, on average the transversal view provides slightly higher scores (1.12). This was expected as avoiding the ribs provides a larger field-of-view of the lungs, and therefore a higher probability of detection of pneumonia-related artifacts. The difference is small, and it does not impact the classification of patients into subgroups for most patients. However, in our case, 4 out of 20 patients would change their subgroup, with 3 of them increasing their subgroup classification with a transversal view and one decreasing its subgroup (Figure 6).

Furthermore, certain pathological findings (B-lines) may have been more represented than others (consolidations and pleural effusions). Despite these limitations, this study represents one of the most controlled studies into the interobserver agreement of LUS findings for COVID-19.

Other studies are possible with the gathered data. For instance, a study of the variability by region (i.e., anterior vs lateral vs posterior), upper and lower, left and right. Furthermore, we did not include AI tools able to evaluate the acquired videos and compare them with human observers. In this work, the AI tool used in [8,16] was not compared, and it will be part of future work.

5. Conclusions

The most reliable LUS findings with COVID-19 were the presence of B lines or determining if a scan is normal. We did not observe statistically significant differences between the longitudinal and transverse scans. The IRR in LUS of COVID-19 patients may benefit from more standardization of the clinical protocols.

Author Contributions: "Conceptualization, J.L.H., J.C. and Y.T.C.; software, C.F., J.L.H., J.C. M.M. R.G.; validation, all authors; formal analysis, C.F., J.L.H, J.C., M.M.; investigation, all authors; resources, J.C., J.L.H., M.M. R.G.; data curation, M.M., C.F.; writing—original draft preparation, J.L.H.; writing—review and editing, all authors; supervision, Y.T.C.; funding acquisition, J.C., J.L.H., Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by CDTI (Spanish acronym: Centre for Industrial Technological Development), funding number COI-20201153. Partially supported by the Google Cloud Research Credits program with the funding number GCP19980904, by the project RTI2018-099118-A-I00 founded by MCIU/AEI/FEDER UE and by the European Commission–NextGenerationEU, through CSIC's Global Health Platform (PTI Salud Global).

Institutional Review Board Statement: The study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Hospital Universitario Puerta de Hierro (protocol code 2.0 5 April 2021 and date of approval April 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Demi, L.; Wolfram, F.; Klersy, C.; De Silvestri, A.; Ferretti, V.V.; Muller, M.; Miller, D.; Feletti, F.; Wełnicki, M.; Buda, N.; et al. New International Guidelines and Consensus on the Use of Lung Ultrasound. *J. Ultrasound Med. n/a*, doi:10.1002/jum.16088.
2. Gil-Rodrigo, A.; Llorens, P.; Luque-Hernández, M.-J.; Martínez-Buendía, C.; Ramos-Rincón, J.-M. Lung Ultrasound Integration in Assessment of Patients with Noncritical COVID-19. *J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* **2021**, *40*, 2203–2212, doi:10.1002/jum.15613.
3. Torres-Macho, J.; Sánchez-Fernández, M.; Arnanz-González, I.; Tung-Chen, Y.; Franco-Moreno, A.I.; Duffort-Falcó, M.; Beltrán-Romero, L.; Rodríguez-Suaréz, S.; Bernabeu-Wittel, M.; Urbano, E.; et al. Prediction Accuracy of Serial Lung Ultrasound in COVID-19 Hospitalized Patients (Pred-Echovid Study). *J. Clin. Med.* **2021**, *10*, 4818, doi:10.3390/jcm10214818.
4. Volpicelli, G.; Gargani, L.; Perlini, S.; Spinelli, S.; Barbieri, G.; Lanotte, A.; Casasola, G.G.; Nogué-Bou, R.; Lamorte, A.; Agricola, E.; et al. Lung Ultrasound for the Early Diagnosis of COVID-19 Pneumonia: An International Multicenter Study. *Intensive Care Med.* **2021**, *47*, 444–454, doi:10.1007/s00134-021-06373-7.
5. Calvo-Cebrián, A.; Alonso-Roca, R.; Rodriguez-Contreras, F.J.; Rodríguez-Pascual, M. de las N.; Calderín-Morales, M. del P. Usefulness of Lung Ultrasound Examinations Performed by Primary Care Physicians in Patients With Suspected COVID-19. *J. Ultrasound Med.* **2021**, *40*, 741–750, doi:10.1002/jum.15444.
6. Ebrahimzadeh, S.; Islam, N.; Dawit, H.; Salameh, J.-P.; Kazi, S.; Fabiano, N.; Treanor, L.; Absi, M.; Ahmad, F.; Rooprai, P.; et al. Thoracic Imaging Tests for the Diagnosis of COVID-19. *Cochrane Database Syst. Rev.* **2022**, *5*, CD013639, doi:10.1002/14651858.CD013639.pub5.
7. Caroselli, C.; Blaivas, M.; Marcosignori, M.; Tung Chen, Y.; Falzetti, S.; Mariz, J.; Fiorentino, R.; Pinto Silva, R.; Gomes Cochicho, J.; Sebastiani, S.; et al. Early Lung Ultrasound Findings in Patients With COVID-19 Pneumonia: A Retrospective Multicenter Study of 479 Patients. *J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* **2022**, *41*, 2547–2556, doi:10.1002/jum.15944.
8. Camacho, J.; Muñoz, M.; Genovés, V.; Herraiz, J.L.; Ortega, I.; Belarra, A.; González, R.; Sánchez, D.; Giacchetta, R.C.; Trueba-Vicente, Á.; et al. Artificial Intelligence and Democratization of the Use of Lung Ultrasound in COVID-19: On the Feasibility of Automatic Calculation of Lung Ultrasound Score. *Int. J. Transl. Med.* **2022**, *2*, 17–25, doi:10.3390/ijtm2010002.
9. Tung-Chen, Y.; Gracia, M.M. de; Díez-Tascón, A.; Alonso-González, R.; Agudo-Fernández, S.; Parra-Gordo, M.L.; Ossaba-Vélez, S.; Rodríguez-Fuertes, P.; Llamas-Fuentes, R. Correlation between Chest Computed Tomography and Lung Ultrasonography in Patients with Coronavirus Disease 2019 (COVID-19). *Ultrasound Med. Biol.* **2020**, *46*, 2918–2926, doi:10.1016/j.ultrasmedbio.2020.07.003.
10. Porcel, J.M. Pleural Diseases and COVID-19: Ubi Fumus, Ibi Ignis. *Eur. Respir. J.* **2020**, *56*, doi:10.1183/13993003.03308-2020.
11. Hernández-Píriz, A.; Tung-Chen, Y.; Jiménez-Virumbrales, D.; Ayala-Larrañaga, I.; Barba-Martín, R.; Canora-Lebrato, J.; Zapatero-Gaviria, A.; Casasola-Sánchez, G.G.D. Importance of Lung Ultrasound Follow-Up in Patients Who Had Recovered from Coronavirus Disease 2019: Results from a Prospective Study. *J. Clin. Med.* **2021**, *10*, 3196, doi:10.3390/jcm10143196.
12. Tung-Chen, Y.; Gil-Rodrigo, A.; Algora-Martín, A.; Llamas-Fuentes, R.; Rodríguez-Fuertes, P.; Marín-Baselga, R.; Alonso-Martínez, B.; Sanz Rodríguez, E.; Llorens Soriano, P.; Ramos-Rincón, J.-M. The lung ultrasound "Rule of 7" in the prognosis of

doi:10.20944/preprints202211.0285.v1

COVID-19 patients: Results from a prospective multicentric study. *Med. Clínica* **2022**, *159*, 19–26, doi:10.1016/j.medcli.2021.07.012.

13. Hussain, A.; Via, G.; Melniker, L.; Goffi, A.; Tavazzi, G.; Neri, L.; Villen, T.; Hoppmann, R.; Mojoli, F.; Noble, V.; et al. Multi-Organ Point-of-Care Ultrasound for COVID-19 (PoCUS4COVID): International Expert Consensus. *Crit. Care Lond. Engl.* **2020**, *24*, 702, doi:10.1186/s13054-020-03369-5.
14. Kumar, A.; Weng, Y.; Graglia, S.; Chung, S.; Duanmu, Y.; Lalani, F.; Gandhi, K.; Lobo, V.; Jensen, T.; Nahn, J.; et al. Interobserver Agreement of Lung Ultrasound Findings of COVID-19. *J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* **2021**, *40*, 2369–2376, doi:10.1002/jum.15620.
15. DeSanti, R.L.; Cowan, E.A.; Kory, P.D.; Lasarev, M.R.; Schmidt, J.; Al-Subu, A.M. The Inter-Rater Reliability of Pediatric Point-of-Care Lung Ultrasound Interpretation in Children With Acute Respiratory Failure. *J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* **2022**, *41*, 1159–1167, doi:10.1002/jum.15805.
16. Fatima, N.; Mento, F.; Zanforlin, A.; Smargiassi, A.; Torri, E.; Perrone, T.; Demi, L. Human-to-AI Interrater Agreement for Lung Ultrasound Scoring in COVID-19 Patients. *J. Ultrasound Med. Off. J. Am. Inst. Ultrasound Med.* **2022**, doi:10.1002/jum.16052.
17. Šustić, A.; Mirošević, M.; Szuldrzynski, K.; Marčun, R.; Haznadar, M.; Podbegar, M.; Protić, A. Inter-Observer Reliability for Different Point-of-Care Lung Ultrasound Findings in Mechanically Ventilated Critically Ill COVID-19 Patients. *J. Clin. Monit. Comput.* **2022**, *36*, 279–281, doi:10.1007/s10877-021-00726-9.
18. Tung-Chen, Y.; Ossaba-Vélez, S.; Acosta Velásquez, K.S.; Parra-Gordo, M.L.; Díez-Tascón, A.; Villén-Villegas, T.; Montero-Hernández, E.; Gutiérrez-Villanueva, A.; Trueba-Vicente, Á.; Arenas-Berenguer, I.; et al. The Impact of Different Lung Ultrasound Protocols in the Assessment of Lung Lesions in COVID-19 Patients: Is There an Ideal Lung Ultrasound Protocol? *J. Ultrasound* **2021**, *25*, 483–491, doi:10.1007/s40477-021-00610-x.
19. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Medica* **2012**, *22*, 276–282.
20. Stemler, S. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Pract. Assess. Res. Eval.* **2019**, *9*, doi:https://doi.org/10.7275/96jp-xz07.
21. Stemler, S. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Pract. Assess. Res. Eval.* **2019**, *9*, doi:https://doi.org/10.7275/96jp-xz07.