

1

1 Word count text: 3032

2 Word count abstract: 291

3 **Artificial Intelligence Neural Network Consistently Interprets Lung Ultrasound**

4 **Artifacts in Hospitalized Patients: A Prospective Observational Study**

5 Running Title: "AI INTERPRETS LUNG US ARTIFACTS"

6 Thomas H. Fox MD¹, Gautam R. Gare², Laura E. Hutchins MD¹, Victor S. Perez MD¹, Ricardo

7 Rodriguez MD, David L. Smith MD³, Francisco X. Brito-Encarnacion MD³, Raman Danrad MD³, Hai V.

8 Tran MD¹, Peter B. Lowery MD¹, David J. Montgomery MD¹, Kevin A. Zamorra MD¹, Amita Krishnan

9 MD¹, John M. Galeotti PhD², Bennett P. deBoisblanc MD¹

10

11 1 Department of Pulmonary/Critical Care, Louisiana State University School of Medicine

12 2 Robotics Institute, Carnegie Mellon University

13 3 Department of Radiology, Louisiana State University School of Medicine

14

15 Please direct all correspondence to Dr. Bennett P. deBoisblanc by email at BDeBoi@lsuhsc.edu.

16

17

18

19

20

21

22

23

24

25

26

27

28

29 **Abstract**

30 **Background:**

31 Interpretation of lung ultrasound artifacts by clinicians can be inconsistent. Artificial intelligence (AI)
32 may perform this task more consistently.

33 **Research Question**

34 Can AI characterize lung ultrasound artifacts similarly to humans, and can AI interpretation be
35 corroborated by clinical data?

36 **Study Design and Methods:**

37 Lung sonograms (n=665) from a convenience sample of 172 subjects were prospectively obtained using a
38 pre-specified protocol and matched to clinical and radiographic data. Three investigators scored
39 sonograms for A-lines and B-lines. AI was trained using 142 subjects and then tested on a separate dataset
40 of 30 patients. Three radiologists scored similar anatomic regions of contemporary radiographs for
41 interstitial and alveolar infiltrates to corroborate sonographic findings. The ratio of oxyhemoglobin
42 saturation:fraction of inspired oxygen (S/F) was also used for comparison. The primary outcome was the
43 intraclass correlation coefficient (ICC) between the median investigator scoring of artifacts and AI
44 interpretation.

45 **Results:**

46 In the test set, the correlation between the median investigator score and the AI score was moderate to
47 good for A lines (ICC 0.73, 95% CI [0.53-0.89]), and moderate for B lines (ICC 0.66, 95% CI [0.55-
48 0.75]). The degree of variability between the AI score and the median investigator score for each video
49 was similar to the variability between each investigator's score and the median score. The correlation
50 among radiologists was moderate (ICC 0.59, 95% CI [0.52-0.82]) for interstitial infiltrates and poor for
51 alveolar infiltrates (ICC 0.33, 95% CI [0.07-0.58]). There was a statistically significant correlation
52 between AI scored B-lines and the degree of interstitial opacities for five of six lung zones. Neither AI
53 nor human-scored artifacts were consistently associated with S/F.

54 **Interpretation:**

55 Using a limited dataset, we showed that AI can interpret lung ultrasound A-lines and B-lines in a fashion
56 that could be clinically useful.

57 **Keywords:**

58 “lung ultrasound”, “artificial intelligence”, “inter-rater variability”

59 **Abbreviations:**

60 AI: Artificial intelligence neural network

61 BLUE 1: Point 1 on left side of thorax

62 BLUE 2: Point 2 on left side of thorax

63 BLUE 3: Point 3 on left side of thorax

64 BLUE 4: Point 1 on right side of thorax

65 BLUE 5: Point 2 on right side of thorax

66 BLUE 6: Point 3 on right side of thorax

67 COPD: Chronic Obstructive Pulmonary Disease

68 ICC: Intraclass correlation coefficient

69 US: Ultrasound

70 S/F: Oxyhemoglobin saturation divided by fraction of inspired oxygen

71

72

73

74

75

76

77

78

79 **Introduction:**

80 Lung ultrasound (US) is used for real time identification and prognostication of lung pathology by
81 clinicians at the bedside. Because of its ease of use and because it does not expose patients to ionizing
82 radiation, this imaging modality has undergone explosive growth in intensive care units, emergency
83 departments, and hospital wards to inform management decisions[1]. Lung ultrasound outperforms
84 traditional radiographs in the diagnosis of some common lung pathologies such as cardiogenic pulmonary
85 edema and pneumothorax[2,3].

86
87 Normally aerated lung attenuates the transmission of sound waves making it difficult to directly visualize
88 disease pathology. Instead, the accurate interpretation of lung ultrasound relies on the characterization of
89 reverberation artifacts that are generated at the interface of unaerated and aerated lung.

90
91 Lichtenstein designated over 40 lung ultrasound artifacts in his seminal work on thoracic sonography[4].
92 A-lines and B-lines are the artifacts most readily understood by practicing clinicians[1]. A-lines are
93 generated by reflection of the ultrasound wavefront back and forth *between* the skin and the pleura. A-
94 lines are present in healthy lung and some pathologic conditions, such as pneumothorax and emphysema.
95 In contrast, B-lines are created by reverberations *within* the first millimeter of diseased but aerated lung.
96 B-lines are most commonly seen when there is either interstitial edema or fibrosis. Increasing numbers of
97 B-lines correspond to increasing disease severity[5,6].

98
99 The collage of reverberation artifacts in a lung ultrasound encodes important diagnostic information.
100 However, the interpretation of this collage is subjective and only semi-quantitative leading to
101 inconsistencies in interpretation, even among ultrasound fellowship-trained clinicians[7].

102

103 Because artificial intelligence systems (AI) can handle many more input variables, they have been shown
104 to outperform humans in complex tasks such as the interpretation of mammograms[8]. We therefore
105 hypothesized that an AI network could be trained to decode lung ultrasound artifacts in a fashion similarly
106 to humans. To test this hypothesis we prospectively enrolled patients in a study investigating the
107 correlations among human and AI scoring of lung US. We corroborated AI sonographic findings with
108 selected clinical and radiographic variables. The specific reverberation artifacts chosen were A-lines and
109 B-lines.

110

111 **Materials and Methods:**

112 This was a prospective, observational study conducted on a convenience sample of 172 adult patients
113 admitted to a university-affiliated hospital from January 2021 to February 2022. The protocol was
114 approved by the local institutional review board (LSU IRB#1509). All patients admitted to the study
115 hospital over 18 years of age were eligible to participate. Written informed consent was obtained prior to
116 the performance of any study related procedures.

117

118 **Ultrasound Protocol**

119 Sonographers consisted of a Pulmonary/Critical Care attending, Pulmonary/Critical Care fellow, and five
120 residents. All sonographers received specific training on the technique of obtaining quality lung
121 ultrasounds. Training consisted of 2 hours of independent, directed learning using an accredited lung
122 ultrasound educational product and 6 hours of didactic training involving lung ultrasound acquisition (S1
123 Appendix).

124

125 Patients were scanned with a point-of-care ultrasound system (X-Porte, Fujifilm Sonosite, Bothell, WA),
126 using a linear array probe (HFL38xp/13-6 MHz) and the following presets: depth 6 cm, near field gain
127 0%, far field gain 100%, mechanical index 0.5, tissue index 0.2, tissue harmonics off. Patient's were

128 scanned in the sitting (preferred), or semirecumbent position at three points on each side: the 2nd
129 intercostal space at the mid-clavicular line, 4th intercostal space at the mamillary line, and 5th intercostal
130 space at the posterior axillary line (Figure 1) similar to the BLUE protocol (BLUE points)[9]. The probe
131 was placed in the intercostal space and oriented parallel to the ribs. Six second clips were obtained at each
132 point.

133

134 **Figure 1.** BLUE points of one hemithorax.

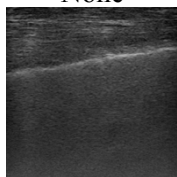
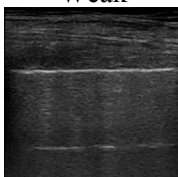
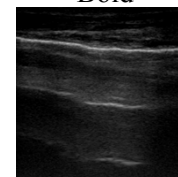
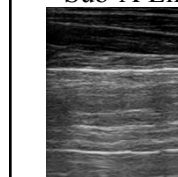
135

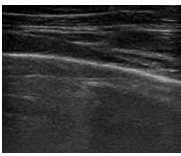
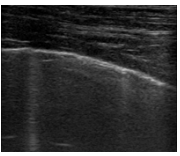
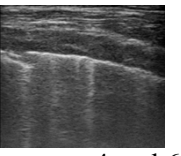

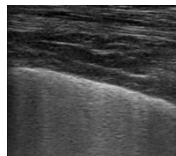
136 **Imaging Interpretation**

137 Sonograms were scored by two physicians and one research staff member (BD, TF, GG), each blinded to
138 clinical data. These investigators received 6 hours of explicit training on lung ultrasound interpretation
139 using accredited training material (S1 Appendix)[10] and had at least an additional 2 years of research
140 experience in this subject matter. Each sonogram was scored for the presence and character of A-lines and
141 the quantity of B-lines on an ordinal scale (Table 1).

142

143 **Table 1.** Lung Ultrasound Artifact Scoring

Artifact	Scoring categories				
A-Lines	None  No A-line	Weak  Faint A-line(s)	Bold  A-line(s) immediately recognizable	Sub-A Lines  A-line(s) immediately recognizable and reflections of fascial planes also identifiable in the subpleural space.	
B-Lines	None	Few (1-3)	Some (4-5)	Many/Coalescing	White Lung

Artifact	Scoring categories				
	 <p>No B-lines</p>	 <p>Between 1 and 3 B-lines</p>	 <p>Between 4 and 6 B-lines</p>	<p>(>6)</p>  <p>More than 6 B-lines, or so many B-lines that individual vertical artifacts cannot be distinguished</p>	 <p>The entirety of the subpleural space is hyperechoic with coalescing B-lines</p>

144

145 Three staff radiologists blinded to the clinical data independently scored the digital chest radiograph in
 146 the test set closest in time to the ultrasound exam. Radiographs obtained less than 24 hours from
 147 ultrasound acquisition were included. Radiographs were scored for the degree of interstitial and alveolar
 148 opacities at 6 different lung zones similar anatomically to the BLUE points (Figure 2)[11]. Each
 149 radiographic sextant was scored on a scale of 0 (no infiltrate) to 3 (dense infiltrates) for both interstitial
 150 and alveolar infiltrates using an electronic slider to provide a continuous variable[11]. To improve
 151 reproducibility, prior to scoring, each radiologist had obtained explicit instruction and had trained using
 152 the scoring system on a separate data set.

153

154 **Figure 2.** Chest radiograph scoring: Each chest radiograph was divided into 6 anatomical zones
 155 corresponding to the 6 ultrasound BLUE points. Two lines separate the thorax in the transverse plane, and
 156 the spinous process divides it sagittally to form six lung zones. Line A is drawn at the level of the inferior
 157 wall of the aortic arch. Line B is drawn at the level of the inferior wall of the right inferior pulmonary
 158 vein.

159

160

161 **Artificial Intelligence Network**

162 A previously published artificial intelligence neural network which has been used to analyze lung
163 ultrasound artifacts[12,13] by employing a Temporal Shift Module (TSM)[14] was trained using 485
164 ultrasound clips from 142 research subjects. The previously published model characterized A and B line
165 artifacts with an accuracy of 76.4% and a precision of 70.8%[13]. The TSM model is video-based model
166 that jointly analyzes a group of frames belonging to a video clip in order to simultaneously predict A-lines
167 and B-lines[14]. Such a video-model is better suited to detect transitory features (B-lines) rather than
168 frame-based models[15,16] that only use a single frame for analysis.

169
170 In the present study, no crossover existed between patients in the training set and those in the test set.
171 Subjects in the training and test sets had sonograms conducted in a similar fashion to that outlined as
172 above. For AI training, each sonogram clip was pre-labeled by one of the investigators (BD, TF, or GG)
173 using an annotator that captured the predominant artifacts. The training involved exposing AI to the
174 labeled video clip, as has been described previously[14].

175
176 Once trained, AI was tasked with interpreting a separate test set of 180 unlabelled clips from 30 patients
177 for A-lines and B-lines. For each clip, AI predicted a probability that selected A-line patterns and selected
178 B-line patterns would be present. Any A-line or B-line pattern with a probability greater than 50% was
179 scored as being present. For example, in a single clip, AI may produce a probability that weak A-lines
180 were present (40%), bold A-lines were present (60%), etc. If two or more A-line patterns were scored as
181 being present, the bolder descriptor for the pattern was chosen. For example, if weak and bold A-lines
182 were scored as being present (had probability scores greater than 50%), the bold A-line option was
183 chosen. In another example, if “few” and “many/coalescing” B-lines were scored as being present, the
184 clip was scored as having “many/coalescent” B-lines. This determination was made *a priori*, and is
185 consistent with clinician scoring. It was not anticipated prior to data interpretation that there would be
186 clips where AI was unable to identify a B-line pattern with a probability greater than 50%. In these

187 instances we chose the B-line pattern with the highest probability, even if that probability was less than
188 50%.

189

190 **Clinical Data**

191 In the test set, the following demographic and clinical descriptors were obtained at the time of each exam:
192 age, gender, admission location, arterial oxygen saturation/fraction of inspired oxygen (S/F), BMI, final
193 diagnosis at discharge, and NIH ordinal scale[17]. S/F was determined using pulse oximetry performed
194 concurrently with the lung US. Diagnosis and NIH ordinal scale was established by reviewing the primary
195 team's documentation, contributing lab and imaging studies, and response to treatments. Method of
196 diagnosis for specific conditions is included in the online supplement(S1 Table).

197

198 **Statistical Analysis**

199 The primary outcome was the intraclass correlation coefficient (ICC) between the median investigator
200 artifact interpretation and the AI artifact interpretation. Secondary outcomes included inter-rater reliability
201 of lung artifact interpretation among investigators, and interrater reliability among chest radiograph
202 scoring among reading radiologists, both determined by the intraclass correlation coefficients. All ICCs
203 were calculated using a two-way effects, absolute agreement, and single rater model and reported with
204 95% confidence intervals.

205

206 External validation was achieved through comparison of AI and investigator artifact interpretation with
207 radiographic characteristics of pulmonary disease, as well as oxygenation as measured by S/F ratio. These
208 relationships were quantified with an analysis of variation (ANOVA). Data were reported as p-values and
209 effect quantification via η^2 with statistical significance defined as a p-value less than 0.05. Prior to
210 conducting the ANOVA, Levene's test of equality was used to confirm homoscedasticity. Relationships
211 between the ultrasound artifact interpretation and the clinical data were visualized using box plots. All
212 analyses were conducted using R version 4.1.3 (R Core Team for Statistical Computing, Austria).

213

214 **Results**

215 Clinical data for research subjects in the test set are shown in Table 2. The average patient age was 66
216 years; the majority were admitted to an ICU (40%); and there was a relatively equal mix of men (53%)
217 and women (47%). The most common diagnoses were decompensated heart failure (n=7) and COVID-19
218 pneumonia (n=7), with bacterial pneumonia, chronic obstructive lung disease (COPD) exacerbations,
219 pleural effusion, and interstitial lung disease making up a minority of diagnoses (Table 3).

220 **Table 2.** Demographic and Clinical Information

Characteristic	N=30 ^a
Age (years)	66 (51,78)
Gender	
Female	14 (47%)
Male	16 (53%)
Disposition	
Inpatient Ward	12 (40%)
Intensive Care Unit	18 (60%)
BMI	28 (26, 32)
S/F	330 (240, 372)
NIH Ordinal Scale	
3	2 (6.7%)
4	2 (6.7%)

5	17 (57%)
6	7 (23%)
7	2 (6.7%)

221 a:Median (25%, 75%); n (%)

222 **Table 3:** Frequency table for diagnosis for enrolled patients in the test set.

Diagnosis*	Number of Patients
Heart Failure Exacerbation	7
COVID-19 Pneumonia	7
Pleural Effusion	5
Interstitial Lung Disease	5
COPD Exacerbation	4
Bacterial Pneumonia	2
Other	2

223 * Multiple diagnoses could exist in the same patient

224

225 **Inter-rater agreements**

226 Among clinicians there was moderate to good agreement overall in A-line pattern description (ICC= 0.75

227 [95% CI: 0.64-0.83]), and moderate agreement in B-line pattern description (ICC= 0.71 [95% CI 0.58-

228 0.79]). AI scoring of A-lines had moderate to good agreement with the median human A-line score using

229 intraclass correlation coefficient (ICC= 0.73 [95% CI 0.53-0.84]). AI scoring of B-lines also showed

230 moderate agreement with median human scoring (ICC= 0.66 [95% CI 0.55-0.75])(Tables 4, 5).

231

232 **Table 4.** ICC among human and AI scoring of A-lines [95% CI]

233

Anatomic Location	ICC among Human Scoring	ICC between AI and Median Human Scoring
BLUE 1 (n=30)	0.92 [0.85-0.96]	0.88 [0.76-0.94]
BLUE 2 (n=30)	0.70 [0.53-0.83]	0.82 [0.66-0.91]
BLUE 3 (n=30)	0.63 [0.36-0.80]	0.62 [0.1-0.85]
BLUE 4 (n=30)	0.74 [0.58-0.85]	0.82 [0.63-0.91]
BLUE 5 (n=30)	0.80 [0.63-0.90]	0.68 [0.43-0.83]
BLUE 6 (n=30)	0.52 [0.24-0.73]	0.43 [0.02-0.70]
Mean (n=180)	0.75 [0.64-0.83]	0.73 [0.53-0.84]

234

235

236 **Table 5.** ICC among AI and median human US scoring of B-lines [95% CI]

Anatomic Location	ICC among Human Scoring	ICC between AI and Median Human Scoring
BLUE 1 (n=30)	0.82 [0.69-0.9]	0.79 [0.48-0.91]
BLUE 2 (n=30)	0.64 [0.4-0.8]	0.62 [0.35-0.80]
BLUE 3 (n=30)	0.65 [0.43-0.81]	0.76 [0.4-0.90]

BLUE 4 (n=30)	0.67 [0.45-0.82]	0.74 [0.5-0.87]
BLUE 5 (n=30)	0.72 [0.53-0.85]	0.50 [0.17-0.72]
BLUE 6 (n=30)	0.77 [0.53-0.89]	0.59 [0.261-0.79]
Mean (n=180)	0.71 [0.58-0.79]	0.66 [0.55-0.75]

237
 238 To more directly compare the variability between AI and each investigator, AI was considered as a
 239 separate investigator. Then for each artifact pattern, the variability between AI and the median score of all
 240 investigators was compared to the variability of each investigator and the overall median. For A-lines, the
 241 ICC between each investigator and the median score for each clip ranged from 0.74 (TF vs median) to
 242 0.83 (BD vs median). AI had similar variability with an ICC of 0.74 when compared to the median
 243 investigator score. For B-lines, investigator variability ranged from 0.6 (ICC between TF vs median of
 244 AI and the other investigators) to 0.75 (ICC between GG vs median of AI and the other investigators)
 245 while AI variability was 0.65 (ICC between AI vs median investigator score) (table 6). Thus AI performed
 246 within the range of human scoring for the detection of specific A-line and B-line artifact patterns.

247
 248 **Table 6.** Variance shown as ICC [95% CI] between Human and AI scoring of Artifacts versus Median

	TF vs median	BD vs median	GG vs median	AI vs median
A-Lines	0.74 [0.48-0.85]	0.83 [0.76-0.88]	0.77 [0.6-0.85]	0.74 [0.53-0.84]
B-Lines	0.6 [0.26-0.77]	0.71 [0.61-0.78]	0.75 [0.64-0.83]	0.65 [0.54-0.73]

249
 250
 251 Furthermore, there was not a significant difference in interrater reliability in any one disease state over
 252 another, although AI scoring was most similar to humans in scoring sonograms of patients with COVID
 253 and least similar in patients with COPD (S2 Table). Although many clinicians use radiology to inform

254 patient care of respiratory diseases, the ICC among radiologist scoring of both interstitial and alveolar
255 infiltrates was moderate to poor (Table 7).

256

257 **Table 7.** Inter-rater reliability among radiologists for interstitial and alveolar opacities shown as ICC
258 [95% CI]

Anatomic Location	Interstitial Scoring	Alveolar Scoring
BLUE 1 (n=30)	0.30 [0.09-0.53]	0.24 [0.04- 0.47]
BLUE 2 (n=30)	0.65 [0.47-0.79]	0.40 [0.18-0.61]
BLUE 3 (n=30)	0.56 [0.35 -0.73]	0.24 [0.04-0.47]
BLUE 4 (n=30)	0.60 [0.4-0.76]	0.00 [-0.16-0.23]
BLUE 5 (n=30)	0.69 [0.52-0.82]	0.35 [0.13-0.57]
BLUE 6 (n=30)	0.44 [0.22-0.65]	0.36 [0.10-0.60]
Mean (n=180)	0.59 [0.52-0.82]	0.329 [0.07-0.58]

259

260

261

262 **Comparisons with Clinical Data**

263 A statistically significant association ($p < 0.05$) was found between the density of interstitial opacities in
264 corresponding chest radiographs and number of B-lines counted by AI in five of six anatomic lung zones
265 using an ANOVA with a large effect size (η^2 range: 0.21-0.47). Similarly, the density of interstitial
266 opacities was associated with investigator-scored B-lines (S4,5 Tables). No statistically significant
267 association was found between the density of interstitial opacities and the strength of A-lines as scored by

268 either AI or investigators using an ANOVA (S6,7 Tables). Box plots of B-lines versus interstitial
269 opacities demonstrated an increased B-line number by both human and AI scoring in lung zones with
270 denser interstitial markings on chest radiographs (Figure 3).

271
272 **Figure 3.** Box plots of radiographic interstitial score and artifact interpretation are shown. Median
273 interstitial score determined by radiologists are shown on the Y-axis. Videos scored for the character of
274 the A-line or number of B-lines by either AI or the median human score. The black bar indicated the
275 median radiographic interstitial scoring. The colored bar represents the 25th and 75th percentile. The
276 extent of the whiskers indicate the 95% confidence interval.

277
278 Three of six lung zones had a statistically significant association ($p < 0.05$) between the degree of alveolar
279 opacities and AI-scored A-lines, with a large effect size (η^2 range: 0.25-0.29). In comparison, two of six
280 lung zones showed a statistically significant association between the degree of alveolar opacities on chest
281 radiographs and investigator-scored A-lines (S9,10 Tables). No statistically significant association was
282 found between the density of alveolar opacities and the number of B-lines on sonograms scored by either
283 AI or investigators (S11,12 Tables).

284
285 A statistically significant association was found between oxygenation via S/F and AI-scored A-lines in
286 three of six lung zones using an ANOVA (S13,14 Tables), and also demonstrated visually by box plots
287 (Figure 4). There was no statistically significant association between S/F and the brightness of A-lines as
288 scored by humans (S15,16 Tables).

289
290 **Figure 4.** Box plots of lung function quantified by S/F and artifact interpretation are shown. S/F for each
291 research subject is shown on the Y axis. Human and AI scoring of A-lines and B-lines are shown. The
292 black bar indicated the median S/F for all videos scored as having each character of A-line or number of

293 B-lines. The colored bars represent the 25th and 75th percentiles. The whiskers indicate the 95%
294 confidence interval.

295

296 **Discussion**

297 Human sonographers show significant variability in the scoring of lung ultrasound artifacts. In spite of
298 this unwanted scoring heterogeneity, point-of-care ultrasound is commonly used to inform patient care
299 decisions. We also observed variability in our human scoring of lung ultrasound artifacts, furthermore the
300 degree of variability was in line with existing evidence on this topic[7,18].

301

302 Unlike human scoring, a fully trained AI network holds the promise of yielding highly reproducible
303 results. In the present study, we observed a moderate correlation between AI and investigator
304 interpretation of A-lines, indicating that AI interpreted clips similarly to investigators for this artifact.
305 There was a weaker correlation between AI and investigator scoring of B-lines, although the degree of
306 correlation was in line with existing evidence on this topic[16,19]. In composite, these data suggest that
307 AI trained on a relatively small data set can interpret A-line and B-line artifacts within the range of human
308 interpretation.

309

310 Chest radiographs are a commonly ordered imaging modality in hospitalized patients. It is clear, however,
311 that like other imaging modalities, there is significant variability in how chest radiographs are
312 interpreted[20]. When compared to the degree of variability among radiologists interpreting interstitial
313 and alveolar infiltrates, human sonographers and AI scored A-line and B-line artifacts with lower
314 interrater variability.

315

316 Previous studies have shown a positive correlation between increasing interstitial infiltrates and the
317 number of B-lines[5,6]. We observed a similar relationship between the density of interstitial infiltrates
318 and the number of B-lines scored by both humans and AI. It is notable that AI did not score any clips in
319 the test set as having the highest B-line severity (3 or “white lung”), perhaps because too few clips of this
320 severity were included in the training set. However, ultrasound clips scored by clinicians as having the
321 highest severity B-line score had wide confidence intervals, which may indicate that this ultrasound
322 finding is not a reliable indicator of worsening interstitial disease.

323

324 It was less clear how alveolar opacities on chest radiographs would correlate with lung ultrasound
325 interpretation. In this dataset, the degree of alveolar opacification, as adjudicated by radiologists, was
326 inversely correlated with the boldness of A-lines as interpreted by AI in three of six lung zones.
327 Somewhat surprisingly, the B-line artifact was not a reliable predictor of alveolar infiltrates on chest
328 radiographs (S12,13 Tables).

329

330 Artificial intelligence neural networks have previously shown the ability to differentiate normal from
331 abnormal lung sonograms, identifying, for example, an A-line predominant versus B-line predominant
332 clip[21]. AI has also been shown to improve novice lung sonographers interpretation[19]. AI systems
333 have previously been able to characterize multiple lung ultrasound artifacts simultaneously compared to a
334 human standard[22]. These studies often do not attempt to analyze AI artifact identification beyond its
335 similarity to human interpretation[23].

336

337 There are two novel aspects to the present study that extend previous observations. First, we tasked AI
338 with characterizing more artifacts in more detail than previous studies. Second, we matched AI ultrasound
339 artifact interpretations not only to human interpretation of the same sonogram, but also to that of an

340 entirely different radiographic modality, as well as to physiologic data. We demonstrated that AI scoring
341 compared favorably to human interpretation of sonograms, and that it correlated with other radiographic
342 data. This provides added clinical relevance to AI interpretation of lung ultrasound artifacts.

343
344 Several protocols have been used to obtain thoracic sonograms for research studies, with variations in
345 probe selection, orientation, and depth settings[24,25]. Of available probes, a high frequency, linear array
346 probe was chosen for the present study because it allows enhanced resolution of the pleural line and
347 subpleural structures[26]. The probe was placed in the intercostal space and oriented parallel to the ribs to
348 allow visualization of a larger area of pleural surface uninterrupted by rib shadow[26,27]. This orientation
349 also allowed for continuous contact with the skin along the entire length of the probe. Six cm was the
350 maximum depth setting available for the probe used in this study.

351

352 **Limitations**

353 Our study has limitations. First, it was conducted on a small convenience sample of patients from a single
354 center, and may not be applicable to broader patient populations with more diverse pathologies. Second,
355 all patients were scanned using an explicit protocol involving a single ultrasound model and probe, which
356 may limit its generalizability to other ultrasound manufacturers, probes, and scanning techniques. Third,
357 there was a limited number of sonographers and human interpreters, and the experience beyond the
358 explicit training received as part of this experiment is not uniform. Fourth, the training set was over-
359 represented with clips taken at the first and fourth BLUE points as opposed to clips more caudally located
360 on the thorax, which may have contributed to the lower reliability in AI rating at these anatomic locations.
361 Fifth, only two artifacts were measured in this study, and some patients' pathology cannot be
362 characterized using these artifacts alone, such as those with pleural effusions. Sixth, we recognize that in
363 very obese patients, there may be more three centimeters of soft tissue between the skin and the pleural
364 line which might have limited our ability to detect A-line artifacts. In this uncommon occurrence, we
365 attempted to compress the probe against the skin until the skin to pleural distance spanned less than 3 cm.

366 Seventh, imaging studies were interpreted by only six investigators (three for ultrasound, three for
367 radiographs), which limits the statistical validity of variability among humans. Despite these limitations,
368 this study represents encouraging evidence of the potential for machine learning to accurately characterize
369 ultrasound artifacts with clinical implications.

370

371 **Conclusions**

372 In this prospective, observational study of a small convenience sample of adults admitted to a university
373 affiliated hospital, we demonstrate that an artificial intelligence network can be trained to identify and
374 characterize A-line and B-line artifacts within the range of variability of human interpreters. We
375 corroborate these interpretations with radiographic and clinical comparators that show AI interpretation of
376 B-lines is associated with degree of interstitial disease.

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391 **Acknowledgements:**

392 BD had full access to all of the data in the study and takes responsibility for the integrity of the data and
393 the accuracy of the data analysis. TF, LH, VP, GG, DS, FB, RD, HT, PL, DM, KZ, and BD acquired data.
394 TF, GG, AK, JG, and BD contributed to the conception and design of the study. TF and GG performed
395 statistical analyses. All authors participated in the interpretation of the data, provided critical feedback
396 and final approval for submission, and took responsibility for the accuracy, completeness, and protocol
397 adherence of data and analyses.

398

399 **References:**

- 400 1) Frankel HL, Kirkpatrick AW, Elbarbary M, et al. Guidelines for the Appropriate Use of
401 Bedside General and Cardiac Ultrasonography in the Evaluation of Critically Ill Patients-
402 Part I: General Ultrasonography. *Crit Care Med*. 2015;43(11):2479-2502.
403 doi:10.1097/CCM.0000000000001216
- 404 2) Maw AM, Hassanin A, Ho PM, et al. Diagnostic Accuracy of Point-of-Care Lung
405 Ultrasonography and Chest Radiography in Adults With Symptoms Suggestive of Acute
406 Decompensated Heart Failure: A Systematic Review and Meta-analysis. *JAMA Netw*
407 *Open*. 2019;2(3):e190703. doi:10.1001/jamanetworkopen.2019.0703
- 408 3) Ding W, Shen Y, Yang J, He X, Zhang M. Diagnosis of pneumothorax by radiography
409 and ultrasonography: a meta-analysis. *Chest*. 2011;140(4):859-866. doi:10.1378/chest.10-
410 2946
- 411 4) Lichtenstein DA. *Whole Body Ultrasonography in the Critically Ill*. Springer Berlin
412 Heidelberg; 2010. doi:10.1007/978-3-642-05328-3
- 413 5) Anile A, Russo J, Castiglione G, Volpicelli G. A simplified lung ultrasound approach to
414 detect increased extravascular lung water in critically ill patients. *Crit Ultrasound J*.
415 2017;9(1):13. doi:10.1186/s13089-017-0068-x

- 416 6) Picano E, Frassi F, Agricola E, Gligorova S, Gargani L, Mottola G. Ultrasound lung
417 comets: a clinically useful sign of extravascular lung water. *J Am Soc Echocardiogr.*
418 2006;19(3):356-363. doi:10.1016/j.echo.2005.05.019
- 419 7) Gullett J, Donnelly JP, Sinert R, et al. Interobserver agreement in the evaluation of B-
420 lines using bedside ultrasound. *J Crit Care.* 2015;30(6):1395-1399.
421 doi:10.1016/j.jcrc.2015.08.021.
- 422 8) McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for
423 breast cancer screening. *Nature.* 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
- 424 9) Lichtenstein DA, Mezière GA. Relevance of lung ultrasound in the diagnosis of acute
425 respiratory failure: the BLUE protocol. *Chest.* 2008;134(1):117-125.
426 doi:10.1378/chest.07-2800
- 427 10) Doerschug KC, Schmidt GA. Intensive care ultrasound: III. Lung and pleural ultrasound
428 for the intensivist. *Ann Am Thorac Soc.* 2013;10(6):708-712.
429 doi:10.1513/AnnalsATS.201308-288OT
- 430 11) Hanley M, Brosnan C, O'Neill D, et al. Modified Brixia chest X-ray severity scoring
431 system and correlation with intubation, non-invasive ventilation and death in a
432 hospitalised COVID-19 cohort. *J Med Imaging Radiat Oncol.* Published online
433 November 29, 2021. doi:10.1111/1754-9485.13361
- 434 12) Gare GR, Tran HV, deBoisblanc BP, Rodriguez RL, Galeotti JM. Weakly Supervised
435 Contrastive Learning for Better Severity Scoring of Lung Ultrasound. arXiv. Published
436 online 2022. doi:10.48550/arXiv.2201.07357
- 437 13) Gare GR, Fox T, Lowery P, et al. Learning Generic Lung Ultrasound Biomarkers for
438 Decoupling Feature Extraction from Downstream Tasks. arXiv. Published online 2022.
439 doi:10.48550/arxiv.2206.08398

- 440 14) Lin, Ji, Chuang Gan and Song Han. "TSM: Temporal Shift Module for Efficient Video
441 Understanding." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*
442 (2019): 7082-7092.
- 443 15) Gare GR, Schoenling A, Philip V, et al. Dense Pixel-Labeling For Reverse-Transfer And
444 Diagnostic Learning On Lung Ultrasound For Covid-19 And Pneumonia Detection. In:
445 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE;
446 2021:1406-1410. doi:10.1109/ISBI48211.2021.9433826
- 447 16) Van Sloun, Ruud JG, and Libertario Demi. "Localizing B-lines in lung ultrasonography
448 by weakly supervised deep learning, in-vivo results." *IEEE journal of biomedical and*
449 *health informatics* 24.4 (2019): 957-964
- 450 17) Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the Treatment of Covid-19 -
451 Final Report. *N Engl J Med.* 2020;383(19):1813-1826. doi:10.1056/NEJMoa2007764
- 452 18) DeSanti RL, Cowan EA, Kory PD, Lasarev MR, Schmidt J, Al-Subu AM. The Inter-
453 Rater Reliability of Pediatric Point-of-Care Lung Ultrasound Interpretation in Children
454 with Acute Respiratory Failure. *J Ultrasound Med.* Published online August 11, 2021.
455 doi:10.1002/jum.15805
- 456 19) Russell FM, Ehrman RR, Barton A, Sarmiento E, Ottenhoff JE, Nti BK. B-line
457 quantification: comparing learners novice to lung ultrasound assisted by machine
458 artificial intelligence technology to expert review. *Ultrasound J.* 2021;13(1):33.
459 doi:10.1186/s13089-021-00234-6
- 460 20) Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver
461 variability in applying a radiographic definition for ARDS. *Chest.* 1999;116(5):1347-
462 1353. doi:10.1378/chest.116.5.1347
- 463 21) Baloescu C, Toporek G, Kim S, et al. Automated Lung Ultrasound B-Line Assessment
464 Using a Deep Learning Algorithm. *IEEE Trans Ultrason Ferroelectr Freq Control.*
465 2020;67(11):2312-2320. doi:10.1109/TUFFC.2020.3002249

- 466 22) Frey, Benjamin, et al. "Multi-stage investigation of deep neural networks for COVID-19
467 B-line feature detection in simulated and in vivo ultrasound images." *Medical Imaging*
468 2022: Computer-Aided Diagnosis. Vol. 12033. SPIE, 2022.
- 469 23) Camacho, Jorge, et al. "Artificial Intelligence and Democratization of the Use of Lung
470 Ultrasound in COVID-19: On the Feasibility of Automatic Calculation of Lung
471 Ultrasound Score." *International Journal of Translational Medicine* 2.1 (2022): 17-25.
- 472 24) Demi L, Wolfram F, Klersy C, et al. New international guidelines and consensus on the
473 use of lung ultrasound. *J Ultrasound Med*. Published online August 22, 2022.
- 474 25) Ball J. Lung ultrasound signs to diagnose and discriminate interstitial syndromes in ICU
475 patients: A diagnostic accuracy study in two cohorts. *Crit Care Med*. 2022;50(11):1678-
476 1680.
- 477 26) Gargani L, Volpicelli G. How I do it: lung ultrasound. *Cardiovasc Ultrasound*.
478 2014;12:25.
- 479 27) von Groote-Bidlingmaier F, Koegelenberg CFN. A practical guide to transthoracic
480 ultrasound. *Breathe*. 2012;9(2):132-42.
- 481
482
483
484
485
486
487
488
489
490
491

24

492

493

494

495

496

497

498

Figure 1. BLUE points of one hemithorax.



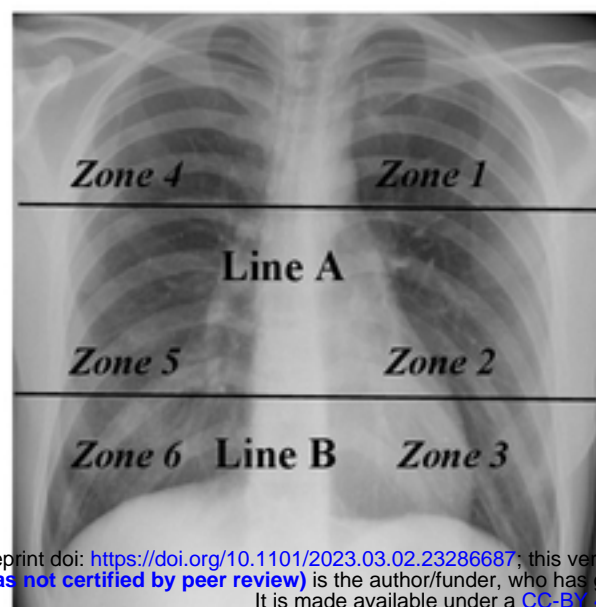
Blue Point 1

Blue Point 2

Blue Point 3

medRxiv preprint doi: <https://doi.org/10.1101/2023.03.02.23286687>; this version posted March 5, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Figure 2. Chest radiograph scoring



medRxiv preprint doi: <https://doi.org/10.1101/2023.03.02.23286687>; this version posted March 5, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#).

Figure 3. Box plots of radiographic interstitial score and artifact interpretation

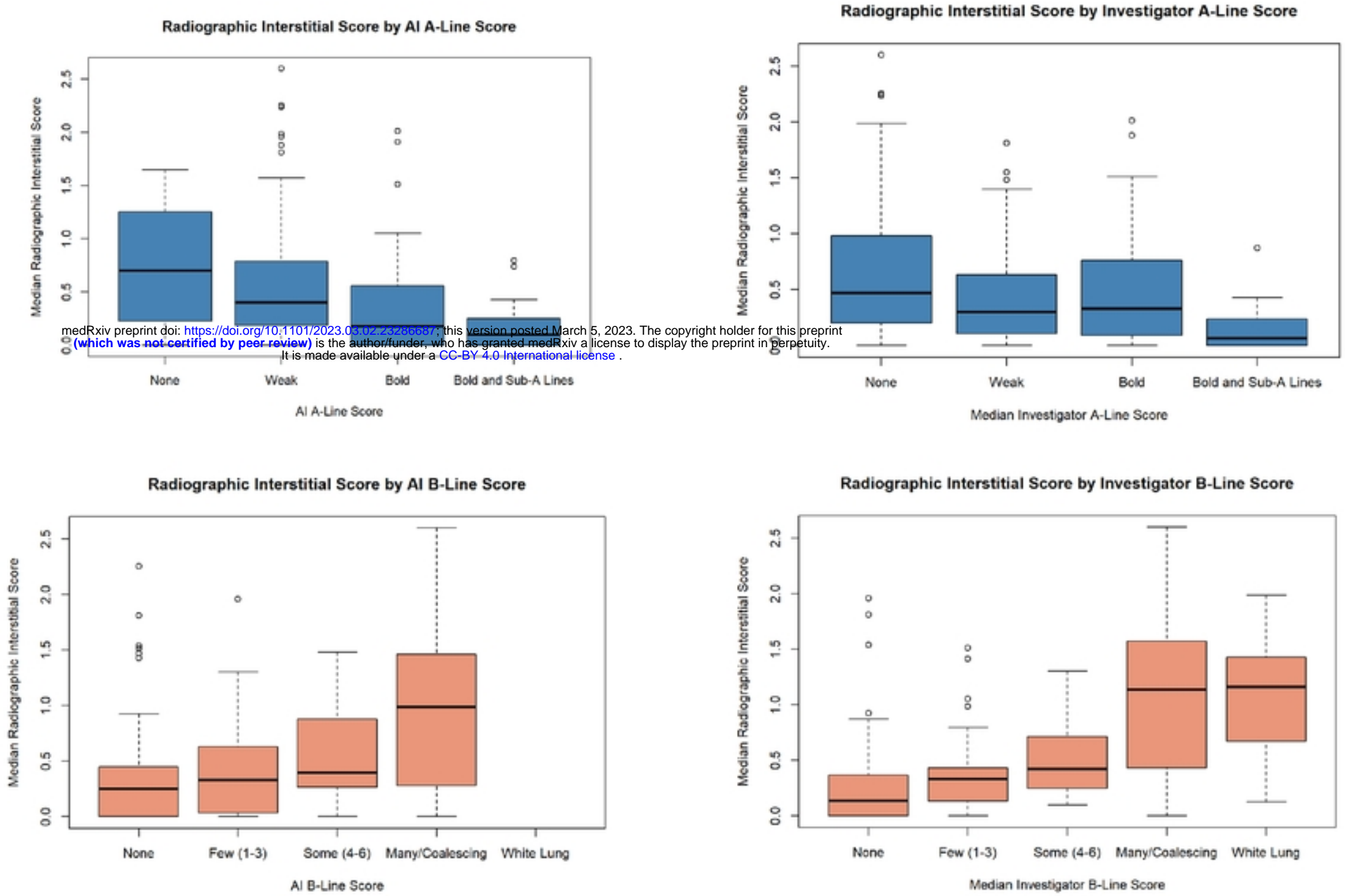


Figure 4. Box plots of lung function quantified by S/F and artifact interpretation

