Received: 28 March 2018 | Revised: 26 April 2018 | Accepted: 7 June 2018

DOI: 10.1111/jvim.15257

Journal of Veterinary Internal Medicine Open Access

ACVIM American College of Veterinary Internal Medicine

STANDARD ARTICLE

# Inter-rater agreement and reliability of thoracic ultrasonographic findings in feedlot calves, with or without naturally occurring bronchopneumonia

S. Buczinski[1] | C. Buathier[1] | A.M. Bélanger[1] | H. Michaux[1] | N. Tison[1] | E. Timsit[2]

[1]Département des Sciences Cliniques, Faculté de Médecine Vétérinaire, Université de Montréal, Québec, Canada

[2]Department of Production Animal Health, Faculty of Veterinary Medicine, University of Calgary, Calgary, Canada

**Correspondence**
S. Buczinski, Dr Vét, MSc, Dip. ACVIM, Département des Sciences Cliniques, Faculté de Médecine Vétérinaire, Université de Montréal, 3200 rue Sicotte, St-Hyacinthe, J2S 2M2, QC, Canada.
Email: s.buczinski@umontreal.ca

**Background:** Thoracic ultrasonography (TUS) can be used to assess the extent and severity of lung lesions associated with bronchopneumonia (BP) in feedlot cattle.

**Hypothesis/Objectives:** To assess inter-rater agreement and reliability of TUS findings in feedlot cattle, with or without naturally occurring BP.

**Animals:** Feedlot steers with (n = 210) or without (n = 107) clinical signs of BP that were assessed by TUS in a previous case-control study.

**Methods:** A random sample of 50 TUS videos (16-s duration) were scored by 6 raters with various levels of TUS expertise. Lung consolidation, comet tail artifacts, pleural irregularity and effusion were scored. Inter-rater agreement was assessed using raw percentage of agreement (Pa), Cohen's and Fleiss' Kappa (κ), and Gwet agreement coefficient (AC1). Intra-class correlation (ICC) was determined for variables with continuous measurements (mixed factorial design).

**Results:** Median (interquartile range [IQR]) Pa were 0.84 (0.80-0.89), 0.82 (0.80-0.87), 0.62 (0.53-0.67), and 0.82 (0.75-0.86) for presence of lung consolidation, comet tails, pleural irregularity, and pleural effusion, respectively. For the same lesions, Fleiss κ (95% confidence intervals [CI]) were 0.67 (0.49-0.86), 0.56 (0.33-0.80), 0.20 (−0.05 to 0.44), and 0.36 (0.10-0.61), respectively. AC1 were 0.68 (0.51-0.86), 0.73 (0.58-0.89), 0.21 (−0.01 to 0.44), and 0.71 (0.51-0.92), respectively. Moderate reliability was found among raters for all quantitative variables (ICC ranged from 0.52 to 0.70).

**Conclusions and Clinical Importance:** Inter-rater agreement was good for presence of lung consolidation, comet tails and pleural effusion (based on Pa and AC1) but was slight to poor for pleural irregularity.

KEYWORDS

bovine respiratory disease, cattle, infectious diseases, radiology and diagnostic imaging, respiratory tract

## 1 | INTRODUCTION

Ante-mortem diagnosis of bronchopneumonia (BP) in feedlot cattle often has poor accuracy.[1,2] In a recent meta-analysis, based on clinical illness, sensitivity and specificity of BP diagnosis were estimated at 0.27 (Bayesian credible intervals [BCI], 0.12-0.65) and 0.92 (BCI, 0.72-0.98), respectively.[2] To improve accuracy of BP diagnosis in feedlots, chute-side tests such as thoracic ultrasonography (TUS) and auscultation can be used to confirm the presence of lung lesions.[1–3]

Thoracic ultrasonography has numerous positive features.[4] It can be performed rapidly chute-side with ultrasound devices commonly used for reproductive imaging.[5] Furthermore, it enables visualization

**Abbreviations:** AC, agreement coefficient; BP, bronchopneumonia; CI, confidence interval; ICC, intra-class correlation coefficient; IQR, interquartile range; Pa, percentage of raw agreement; TUS, thoracic ultrasonography.

and quantification of lung and pleural lesions including lung consolidation, cavitary lesions, comet tail artifacts (also called B-lines), pleural irregularity, pleural effusion, and pneumothorax.[4] Among these lesions, presence and depth of lung consolidations are of major interest, because they were associated with an increased risk of mortality in feedlot cattle with naturally occurring BP.[5]

However, TUS is considered an operator-dependent technology[6] and apparently no data on inter-rater agreement and reliability of TUS findings in feedlot calves with naturally occurring BP have been reported. Such information is crucial to understand limitations of TUS and identify findings for which inter-rater agreement and reliability are high and those for which focused, supervised training is needed to ensure correct TUS interpretation.

The objective of our study was to assess inter-rater agreement and reliability of lung consolidation and other TUS findings in feedlot calves, with or without naturally occurring BP. Raters with a wide range in TUS expertise participated. Our hypothesis was that good agreement and reliability would be present among raters for interpretation of lung consolidation.

## 2 | MATERIALS AND METHODS

This project was designed according to the guidelines for reporting reliability and agreement study (GRAAS).[7]

### 2.1 | Thoracic ultrasound video library

Thoracic ultrasound videos were obtained in a case–control study conducted in newly received beef cattle (body weight [BW] = 574 ± 99 lbs), with (n = 210) and without (n = 107) clinical signs of BP. The cattle population sample has been described previously.[8] Briefly, the study was conducted in high BP risk steers and heifers in western Canada. Cattle with ≥1 BP sign (among nasal and ocular discharge, tachypnea, dyspnea or lethargy), rectal temperature ≥40°C and abnormal lung sounds were defined as cases and pen-matched with control calves (2:1 ratio) with no BP signs, rectal temperature <40°C and no abnormal lung sounds at auscultation performed by a veterinarian. The control calves had no history BP treatment and remain health within 60 days after inclusion. The calves were screened in standard squeeze-chutes. All TUS videos (16-s duration) were obtained by the same operator (Nicolas Tison, rater 6) using an IbexPro (EI Medical Imaging, Loveland, CO) device with a 6.2 MHz linear probe, maximal depth = 9 cm, total gain = 32 dB (far gain = 36 dB, near gain = 13 dB). The clinician performing TUS examinations stored only videos for which a pleural line was observed, excluding videos where, because of cattle movement, ribs were observed too often, precluding off-line evaluation.

### 2.2 | Sample size calculation

In the absence of data on inter-rater agreement and reliability for TUS findings in feedlot cattle, sample size (ie, minimum number of TUS videos) was determined based on the inter-rater Cohen's kappa (κ) values for lung consolidation reported in a study of dairy calves.[9] In that study, inter-rater agreement (κ) ranged from 0.6 to 1.0. Using a freely available software (package irr [Gamer M, Lemon J, Fellows I, Singh P. Package IRR, Various Coefficients of Interrater Reliability and Agreement Version 0.84 https://cran.r-project.org/web/packages/irr/irr.pdf], argument N.cohen.kappa; R [R, version 3.3.3 Core team (2013). R: a language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria, URL http://www.R-project.org/]), various simulations were performed to obtain these values (0.6-1.0, by 0.1 step increment) with a lower bound of 95% CI at 0.4. A sample size of 50 TUS videos was determined as suitable with Type I error (α) set at 5%, Type-II error (β) set at 20% and prevalence of lung consolidation ranging from 0.1 to 0.5 (by 0.1 step increment). Videos are available on-line (Supporting Information Video Files).

### 2.3 | Rater selection

Raters enrolled in the study (n = 6) had various levels of TUS expertise, ranging from beginner to expert (Table 1). The pool of raters included a recent DVM graduate performing an internship in cattle health, 2 active researchers on bovine respiratory disease (1 with recent experience in TUS and the other with extensive experience in TUS) and 3 clinicians routinely conducting genital and extra-genital ultrasonography on cattle. Each rater received 30 min of basic training regarding how to assess TUS videos. This training, presented as a slide-show (Supporting Information File 1), indicated all lesions or items to be reported and how to enter findings in a specific spreadsheet.

### 2.4 | Video selection and assessment

Thoracic ultrasound videos (n = 50) were randomly selected from a database of 402 videos using the RAND function in Excel [Windows, Richmond, WA]. For each video, lung consolidation (defined as depth of consolidation ≥1 cm), comet tail artifacts (also called B-lines), pleural irregularity (defined as non-smooth pleural line), pleural effusion (pleural fluid ≥0.5 cm), and cavitary lesions were reported as dichotomous variables (ie, present versus absent).

Quantitative assessment was performed for maximal depth (cm) and area ($cm^2$) of lung consolidation using the grid line (1 $cm^2$) of

**TABLE 1** Experience of raters used for assessing thoracic ultrasound videoloops

| | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5 | Rater 6 |
|---|---|---|---|---|---|---|
| Working experience (y) | 15 | 11 | 15 | 9 | 0 | 8 |
| Experience genital ultrasound | Average | Average | Advanced | Average | Average | Advanced |
| Experience extra-genital ultrasound | Expert | Average | Expert | Expert | Beginner | Average |
| Experience in BRD research (y) | 10 | 11 | 0 | 0 | 0 | 1 |
| Experience in thoracic ultrasound | Expert | Beginner | Advanced | Advanced | Beginner | Advanced |

the video recording. The maximum number of comet tails visible in a frozen image also was reported, as was maximal depth of pleural fluid (cm; when pleural fluid was observed). Videos were saved in .avi format and viewed on a laptop without using specific software (comparable to a rapid chute-side examination). Operators were able to view videos as many times as required and were able to use a frame-by-frame assessment for completing their spreadsheets. Time required for scoring each video was recorded.

### 2.5 | Statistical analyses

All statistical analyses were performed using commercial software (SAS v9.4, SAS, Cary, NC). Inter-rater agreement for each dichotomous variable was assessed using various indices. The raw percentage of agreement (Pa) was noted as a crude marker of concordant pairs of ratings. The Pa is defined as the total number of examinations where agreement is noted by the 2 raters divided by the total number of loops scored (n = 50). A minimum of 0.75 was defined as an acceptable percentage of agreement.[10,11] Agreement beyond chance was assessed using Cohen's Kappa (κ) test between pairs of raters. The κ reports raw agreement corrected for agreement due to chance (Pc),[12]

$$\kappa = (Pa - Pc)/(1 - Pc). \quad (1)$$

General inter-rater agreement was evaluated using Fleiss' κ for multiple raters and Gwet's agreement coefficient type 1 (AC1). The Fleiss κ represents the average pairwise agreement between raters, averaged over all raters' pairs and specimens. Gwet's AC1 provides a chance-corrected agreement coefficient, in line with the percentage level of agreement.[13,14] This agreement measure is useful for interpreting tests with high raw agreement percentages but low κ values due to κ paradoxes (which can occur when a reported anomaly has low prevalence[15]).

Cohen's and Fleiss' κ and Gwet's AC1 were interpreted using previously reported guidelines[16] as follows: poor agreement for values below <0.20; slight agreement for values between 0.21 and 0.40; moderate agreement for values between 0.41 and 0.60; good agreement for values between 0.61-0.80; and, very good agreement for values between 0.81-1.00.

For quantitative variables, the intra-class correlation coefficient (ICC) was calculated using a mixed factorial design.[14,17] The ICC is an indicator of variance between subjects' measures variance versus all other sources of variability. For the mixed factorial design, the rating or value y attributed to the *i*th animal (or subject [s]) by the *j*th rater (r) is defined as follows:

$$y_{ij} = \mu + s_i + r_j + (sr)_{ij} + e_{ij} \quad (2)$$

where μ is the value of the measurement, $r_j$ is the fixed rater effect assuming:

$$\sum_{j=1}^{6} r_j = 0$$

$s_i$ the subject (animal) random effect; $s_i \sim$ Normal (0; $\sigma_s^2$), $(sr)_{ij}$ is the random subject * rater interaction effect; $(sr)_{ij} \sim$ Normal (0;$\sigma_{sr}^2$). The model also assumes that for any subject i:

$$\sum_{j=1}^{6} (sr)_{ij} = 0$$

Finally, $e_{ij}$ is the random error term; $e_{ij} \sim$ Normal (0;$\sigma_e^2$). Interrater reliability (intra-class correlation coefficient (ICC) or ρ) was defined as follows[14]:

$$\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/(r-1)}{\sigma_s^2 + \sigma_{sr}^2 + \sigma_e^2} \quad (3)$$

The ICC was interpreted using a previously reported guideline[18] as follows: ICC ≤ 0.5 = poor indicator of reliability; 0.5 < ICC ≤ 0.75 = moderate reliability; 0.75 < ICC ≤ 0.9 = good reliability; and >0.9 = excellent reliability.

## 3 | RESULTS

Descriptive results of TUS findings are summarized in Table 2. The dataset is available as Supporting Information File 2. The 50 loops were obtained from 46 different animals (38 animals with BP and 8 control calves). Four animals had 2 different videoloops. The mean (SD) time required to perform a complete examination of a videoloop was 0.9 min (0.1) for rater 1, 4.3 min (0.2) for rater 2, 2.1 min (0.1) for rater 3, 1.6 min (0.1) for rater 4, 1.7 min (0.1) for rater 5, and 1.2 min (0.1) for rater 6. The median Pa (IQR) between pairs of raters was 0.84 (0.80-0.89) for the presence of lung consolidation (using ≥1 cm cut-off ), 0.82 (0.80-0.87) for comet tails, 0.62 (0.53-0.67) for pleural irregularity, 0.82 (0.75-0.86) for pleural fluid accumulation (using ≥0.5 cm cut-off ), and 0.94 (0.84-1.00) for cavitary lesions (Table 3). The median (IQR) Cohen's κ was 0.68 (0.60-0.78) for presence of lung consolidation, 0.54 (0.50-0.66) for comet tails, 0.25 (0.08-0.33) for pleural irregularity and 0.34 (0.21-0.48) for pleural fluid (Table 4). Cohen's κ could only be calculated for cavitary lesions for 3 raters' pairs (raters 3, 4 and 5) and were 0.2 (raters 3–4), 0.56 (raters 3–5) and 0.23 (raters 4–5).

The Fleiss κ and AC1 for multiple raters are shown in Figure 1. The Fleiss κ (95% CI) were 0.67 (0.49-0.86) for lung consolidation, 0.56 (0.33-0.80) for comet tails, 0.20 (−0.05 to 0.44) for pleural irregularity and 0.36 (0.10-0.61) for pleural effusion. The AC1 were 0.68 (0.51-0.86) for lung consolidation, 0.73 (0.58-0.89) for comet tails, 0.21 (−0.01 to 0.44) for pleural irregularity, and 0.71 (0.51-0.92) for pleural effusion.

Intra-class correlation coefficients had moderate reliability values of 0.70 for maximal depth of consolidation (cm), 0.68 for maximal area of consolidation ($cm^2$), 0.60 for maximal number of comet tails by frame (n), and 0.52 for maximal depth of pleural fluid accumulation (cm).

## 4 | DISCUSSION

Although TUS commonly is used in dairy calves for BP diagnosis,[9,19–22] information on TUS findings and interpretation is lacking in feedlot cattle.[1] To our knowledge, the present study is the first to report inter-rater agreement and reliability of TUS findings in feedlot calves. Based on a comparison of TUS findings among

**TABLE 2** Descriptive characteristics of ultrasound abnormalities found by 6 different raters assessing 50 ultrasonographic videoloops of feedlot calves with or without naturally occurring bovine respiratory disease complex

| Proportion of anomalies found (n) | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5 | Rater 6 | Median proportion |
|---|---|---|---|---|---|---|---|
| Comet-tails | 0.22 *(11)* | 0.60 *(15)* | 0.60 *(15)* | 0.20 *(10)* | 0.24 *(12)* | 0.26 *(13)* | 0.25 |
| Pleural irregularity | 0.18 *(9)* | 0.58 *(29)* | 0.62 *(31)* | 0.28 *(14)* | 0.58 *(29)* | 0.44 *(22)* | 0.51 |
| Pleural fluid (≥0.5cm) | 0.16 *(8)* | 0.14 *(7)* | 0.08 *(4)* | 0.34 *(17)* | 0.14 *(7)* | 0.28 *(14)* | 0.15 |
| Lung consolidation (≥1 cm) | 0.56 *(28)* | 0.48 *(24)* | 0.64 *(32)* | 0.64 *(32)* | 0.52 *(26)* | 0.50 *(25)* | 0.54 |
| Cavitary lesions | 0 *(0)* | 0 *(0)* | 0.06 *(3)* | 0.20 *(10)* | 0.14 *(7)* | 0 *(0)* | 0.03 |

operators with a range of expertise, inter-rater agreement was good for presence of lung consolidation, comet tails, and pleural effusion (based on Pa and AC1), but slight to poor for detection of pleural irregularities and cavitary lesions. Moderate reliability was found among raters for all quantitative variables (ie, maximal number of comet tails per frame, maximal depth or area of lung consolidation, and maximal depth of pleural effusion).

Our study had several strengths. First, it was designed and reported according guidelines for reporting reliability and agreement study (GRAAS).[7] Secondly, raters had a wide range of TUS expertise (recent DVM graduate to TUS experts), ensuring good external validity. Finally, in addition to Cohen's κ, various measures of agreement were reported, including Fleiss' κ and AC1. These measures have several advantages over Cohen's κ.[14] The Fleiss' κ represents the average pairwise agreement among raters averaged over all raters' pairs and specimens, in contrast to Cohen's κ that only assesses agreement between 2 raters. Compared with κ, Gwet's AC1 is less influenced by prevalence and thus not impacted by κ paradoxes (ie, low κ despite high raw percentage of agreement) and is more stable (less influenced by table asymmetry and operator bias).[13–15] For example, in our study, because of κ paradoxes, κ agreements were low for pleural effusion despite good Pa, whereas AC1 gave a result similar to Pa. To avoid κ paradoxes, some authors recommend that κ not be reported when the prevalence of a variable is not close to 50%.[23] Furthermore, AC1 adjusts chance agreement based on the fact that subjects were "easy" versus "hard" to classify (which are both latent variables in the dataset). With this approach, chance agreement is considered to occur more frequently in "hard" subjects than in "easy" subjects, which can be considered more clinically relevant.

Raters did not perform TUS examinations, but only interpreted TUS findings based on videos. Therefore, our study assessed inter-rater agreement and reliability for TUS video interpretation and not TUS examination itself. However, in our opinion, interpretation of TUS findings is the main source of disagreement among raters, because only minimal training and skills are required to obtain good TUS videos.[4] Regardless, this approach could have underestimated inter-raters agreement and reliability for TUS findings. Indeed, when conducting TUS examination, the operator can easily retake a video if sonographic images are of poor quality, difficult to interpret or both, which was not an option for raters in our study. Furthermore, the rater also can improve his or her interpretation of TUS findings by knowing the location of the sonographic probe and thus determining if the images observed are associated with the myocardium, diaphragm or any intra-abdominal organ versus a pulmonary or pleural

**TABLE 3** Heat-plot summarizing the raw percentage of agreement between 6 raters assessing thoracic ultrasonographic videoloops in feedlot calves with naturally occurring bovine respiratory disease complex

| Rater | Comet tail[a] | Irregularity | Pleural fluid[b] | Consolidation[c] | Cavitary lesion |
|---|---|---|---|---|---|
| 1_2 | 0.8 | 0.4 | 0.9 | 0.92 | 1 |
| 1_3 | 0.92 | 0.52 | 0.88 | 0.88 | 0.94 |
| 1_4 | 0.86 | 0.62 | 0.82 | 0.8 | 0.8 |
| 1_5 | 0.9 | 0.42 | 0.82 | 0.92 | 0.86 |
| 1_6 | 0.92 | 0.62 | 0.76 | 0.9 | 1 |
| 2_3 | 0.8 | 0.68 | 0.9 | 0.8 | 0.94 |
| 2_4 | 0.74 | 0.62 | 0.72 | 0.76 | 0.8 |
| 2_5 | 0.78 | 0.64 | 0.84 | 0.84 | 0.86 |
| 2_6 | 0.8 | 0.62 | 0.82 | 0.9 | 1 |
| 3_4 | 0.78 | 0.66 | 0.74 | 0.84 | 0.82 |
| 3_5 | 0.82 | 0.72 | 0.9 | 0.84 | 0.92 |
| 3_6 | 0.88 | 0.74 | 0.76 | 0.8 | 0.94 |
| 4_5 | 0.84 | 0.5 | 0.76 | 0.8 | 0.78 |
| 4_6 | 0.86 | 0.68 | 0.66 | 0.74 | 0.8 |
| 5_6 | 0.82 | 0.54 | 0.74 | 0.86 | 0.86 |

Green cells are cells with raw percentage of agreement ≥0.75, yellow cells where agreement is between 0.51 and 0.74, red for cells with raw agreement ≤0.5.

[a] Comet tail was defined as either large B-lines or small comet-tails artifacts.

[b] Pleural fluid detection was considered positive if ≥0.5cm of fluid was detected.

[c] Consolidation was considered positive if ≥3cm lung consolidation was detected.

**TABLE 4** Cohen's kappa agreement between 6 raters assessing thoracic ultrasonographic videoloops in feedlot calves with naturally occurring bovine respiratory disease complex

| Raters | Comet tail[a] | Irregularity | Pleural fluid[b] | Consolidation[c] | Cavitary lesions |
|---|---|---|---|---|---|
| 1_2 | 0.485 | −0.089 | 0.608 | 0.841 | – |
| 1_3 | 0.794 | 0.168 | 0.440 | 0.752 | 0.000 |
| 1_4 | 0.578 | −0.058 | 0.540 | 0.586 | 0.000 |
| 1_5 | 0.718 | 0.057 | 0.295 | 0.839 | 0.000 |
| 1_6 | 0.781 | 0.177 | 0.315 | 0.800 | – |
| 2_3 | 0.524 | 0.334 | 0.494 | 0.604 | 0.000 |
| 2_4 | 0.316 | 0.290 | 0.272 | 0.525 | 0.000 |
| 2_5 | 0.444 | 0.261 | 0.336 | 0.681 | 0.000 |
| 2_6 | 0.505 | 0.254 | 0.473 | 0.800 | – |
| 3_4 | 0.421 | 0.385 | 0.289 | 0.653 | 0.237 |
| 3_5 | 0.546 | 0.418 | 0.494 | 0.676 | 0.563 |
| 3_6 | 0.703 | 0.495 | 0.239 | 0.560 | 0.000 |
| 4_5 | 0.535 | 0.066 | 0.376 | 0.596 | 0.225 |
| 4_6 | 0.607 | 0.324 | 0.209 | 0.480 | 0.000 |
| 5_6 | 0.520 | 0.097 | 0.239 | 0.720 | 0.000 |

Green cells are cells with κ ≥ 0.60, yellow cells where 0.20≤ κ < 0.60, red for cells with κ < 0.20.
[a] Comet tail was defined as either large B-lines or small comet-tails artifacts.
[b] Pleural fluid detection was considered positive if ≥0.5cm of fluid was detected.
[c] Consolidation was considered positive if ≥3cm lung consolidation was detected.

anomaly. On the other hand, we could not completely exclude overestimation due to the fact that all raters scored the same loop. Our study design could not answer this specific question.

Good inter-rater agreement and reliability for detection of lung consolidation in our study were consistent with previous studies conducted in humans. Indeed, very good agreement (κ = 0.83) was found for detection of lung consolidation in an Italian emergency department population.[24] Furthermore, in a recent study on childhood pneumonia in Peru, agreement (κ) of 0.77 (0.75-0.78) was found among general practitioners for detection of lung consolidation. Furthermore, this agreement increased to 0.87 (0.86-0.89) when only medium and large lung consolidations were considered.[25] Interestingly, in that study, inter-rater agreement decreased to 0.38 (0.27-0.41) when only minimal pleural abnormalities and comet tails (ie, interstitial abnormalities) were considered, similar to the findings in our study (ie, slight to poor agreement for pleural irregularities).

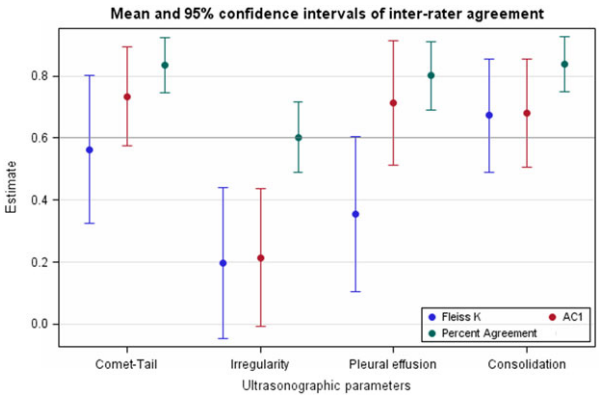


**FIGURE 1** Inter-rater agreement for comet-tail artifacts, pleural irregularity, pleural effusion and consolidation diagnosis between 6 raters scoring 50 video-loops of feedlot calves. Fleiss K: Fleiss Kappa; AC1: Gwet agreement coefficient type 1 for multiple raters. The horizontal line with an estimate at 0.6 was the lower limit for defining a clinically acceptable agreement

Lung consolidation is one of the TUS findings most commonly associated with negative outcome.[4,20,26,27] Therefore, good inter-rater Fleiss' κ and AC1 for lung consolidation should encourage clinicians and researchers to report this variable in future BP studies. However, reliability for maximal depth or area of lung consolidation was only moderate among raters. Therefore, training or software to automatically measure maximum depth and area of lung consolidation is needed to ensure good inter-raters reliability. Good AC1 values, despite slight (for pleural effusion) to moderate (for comet-tails) Fleiss' κ also were encouraging but should further be confirmed in a study with a trait prevalence closer to 50% for avoiding κ paradoxes. Definitions of pleural irregularity vary among raters and therefore are of limited interest. Unfortunately, it was difficult to evaluate agreement for cavitary lesions in our study (due to very low prevalence).

In conclusion, we demonstrated that presence of lung consolidation had good inter-rater agreement. However, reliability of consolidation extension measures (maximal depth and area of consolidation) was only moderate when assessed by multiple raters without specific dedicated software.

## CONFLICT OF INTEREST DECLARATION

Sébastien Buczinski serves as Consulting Editor for Experimental Design and Statistics for the Journal of Veterinary Internal Medicine. He was not involved in review of this manuscript.

## OFF-LABEL ANTIMICROBIAL DECLARATION

Authors declare no off-label use of antimicrobials.

## INSTITUTIONAL ANIMAL CARE AND USE COMMITTEE (IACUC) OR OTHER APPROVAL DECLARATION

The IACUC approval was obtained at the University of Calgary where the ultrasound examinations were performed. Because this study only involved use of ultrasound videoloops previously stored, no animals were used in the study.

## ORCID

*S. Buczinski* http://orcid.org/0000-0002-8460-4885

## REFERENCES

1. Wolfger B,Timsit E,White BJ, et al. A systematic review of bovine respiratory disease diagnosis focused on diagnostic confirmation, early detection, and prediction of unfavorable outcomes in feedlot cattle. *Vet Clin North Am Food Anim Pract*. 2015;31:351–365.
2. Timsit E,Dendukuri N,Schiller I, et al. Diagnostic accuracy of clinical illness for bovine respiratory disease (BRD) diagnosis in beef cattle placed in feedlots: a systematic literature review and hierarchical Bayesian latent-class meta-analysis. *Prev Vet Med*. 2016;135:67–73.
3. Mang AV,Buczinski S,Booker CW, et al. Evaluation of a computer-aided lung auscultation system for diagnosis of bovine respiratory disease in feedlot cattle. *J Vet Intern Med*. 2015;29:1112–1116.
4. Babkine M,Blond L. Ultrasonography of the bovine respiratory system and its practical application. *Vet Clin North Am Food Anim Pract*. 2009;25:633–649.
5. Rademacher RD,Buczinski S,Tripp HM, et al. Systematic thoracic ultrasonography in acute bovine respiratory disease of feedlot steers: impact of lung consolidation on diagnosis and prognosis in a case-control study. *Bov Pract*. 2014;48:1–10.
6. Mayo PH,Beaulieu Y,Doelken P, et al. American College of Chest Physicians/La Societe de Reanimation de Langue Francaise statement on competence in critical care ultrasonography. *Chest*. 2009;135:1050–1060.
7. Kottner J,Audige L,Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96–106.
8. Timsit E,Hallewell J,Booker C, et al. Prevalence and antimicrobial susceptibility of *Mannheimia haemolytica*, *Pasteurella multocida*, and *Histophilus somni* isolated from the lower respiratory tract of healthy feedlot cattle and those diagnosed with bovine respiratory disease. *Vet Microbiol*. 2017;208:118–125.
9. Buczinski S,Forté G,Bélanger AM. Short communication: ultrasonographic assessment of the thorax as a fast technique to assess pulmonary lesions in dairy calves with bovine respiratory disease. *J Dairy Sci*. 2013;96:4523–4528.
10. Burn CC,Weir AA. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. *Vet J*. 2011;188:166–170.
11. Buczinski S,Faure C,Jolivet S, et al. Evaluation of inter-observer agreement when using a clinical respiratory scoring system in pre-weaned dairy calves. *N Z Vet J*. 2016;64:243–247.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37.
13. Wongpakaran N,Wongpakaran T,Wedding D, et al. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:61.
14. Gwet KL. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD: Advanced Analytics, LLC; 2014.
15. Walsh P,Thornton J,Asato J, et al. Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ*. 2014;2:e651.
16. Altman DG. *Practical Statistics for Medical Research*. Boca Raton: Chapman & Hall/CRC; 2006.
17. Shrout PE,Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
18. Koo TK,Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
19. Ollivett TL,Caswell JL,Nydam DV, et al. thoracic ultrasonography and bronchoalveolar lavage fluid analysis in holstein calves with subclinical lung lesions. *J Vet Intern Med*. 2015;29:1728–1734.
20. Ollivett TL,Buczinski S. On-Farm Use of ultrasonography for bovine respiratory disease. *Vet Clin North Am Food Anim Pract*. 2016;32:19–35.
21. Love WJ,Lehenbauer TW,Van Eenennaam AL, et al. Sensitivity and specificity of on-farm scoring systems and nasal culture to detect bovine respiratory disease complex in preweaned dairy calves. *J Vet Diagn Invest*. 2016;28:119–128.
22. Teixeira AG,McArt JA,Bicalho RC. Thoracic ultrasound assessment of lung consolidation at weaning in Holstein dairy heifers: reproductive performance and survival. *J Dairy Sci*. 2017;100:2985–2991.
23. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J Clin Epidemiol*. 2000;53:499–503.
24. 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.
25. Ellington LE,Gilman RH,Chavez MA, et al. Lung ultrasound as a diagnostic tool for radiographically-confirmed pneumonia in low resource settings. *Respir Med*. 2017;128:57–64.
26. Rademacher RMBS,Edmonds M,Tripp HT,Johnson E. Systematic thoracic ultrasonography in acute bovine respiratory disease of feedlot steers: impact of lung consolidation on diagnosis and prognosis in a case-control study. *Bov Prac*. 2014;41:1–10.
27. Rabeling B,Rehage J,Dopfer D, et al. Ultrasonographic findings in calves with respiratory disease. *Vet Rec*. 1998;143:468–471.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Buczinski S, Buathier C, Bélanger AM, Michaux H, Tison N, Timsit E. Inter-rater agreement and reliability of thoracic ultrasonographic findings in feedlot calves, with or without naturally occurring bronchopneumonia. *J Vet Intern Med*. 2018;32:1787–1792. https://doi.org/10.1111/jvim.15257