Journal of Veterinary Internal Medicine

Journal of Veterinary Internal Medicine Open Access
ACVIM American College of Veterinary Internal Medicine

STANDARD ARTICLE OPEN ACCESS

Food & Fiber Animal Respiratory

Agreement of Specific Lung Sounds Auscultation by Veterinarians for the Detection of Bronchopneumonia in Calves

Leticia Princisval[1,2] | Antonio Boccardo[3] | Davide Pravettoni[3] | Salvatore Ferraro[4] | Jean-François Valarcher[4] | Viviani Gomes[5] | Gilles Fecteau[1] | Sébastien Buczinski[1]

[1]Département Des Sciences Cliniques, Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, Québec, Canada | [2]Faculdade de Medicina Veterinaria, Fluminense Federal University, Niterói, Brazil | [3]Dipartimento di Medicina Veterinaria e Scienze Animali (DIVAS), Università Degli Studi di Milano, Lodi, Italy | [4]Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden | [5]Department of Internal Medicine, College of Veterinary Medicine and Animal Science, University of São Paulo, São Paulo, Brazil

Correspondence: Sébastien Buczinski (s.buczinski@umontreal.ca)

Received: 14 March 2025 | **Revised:** 23 July 2025 | **Accepted:** 25 July 2025

Funding: The authors received no specific funding for this work.

Keywords: bovine respiratory disease | lung sound | pneumology | reliability

ABSTRACT

Background: Lung auscultation is a common method for the routine diagnosis of calf bronchopneumonia. However, its repeatability among operators has been criticized.

Objective: Determine agreement among veterinarians for specific lung sounds after a short tutorial to standardize the definition of lung sounds.

Animals: Forty lung sounds from a larger dataset collected at 4 veal calf farms that housed 495–815 animals were submitted online to 10 different veterinarians.

Methods: After a short tutorial on lung sound auscultation, the raters were asked to detect the presence of any abnormal sounds and to differentiate among wheezes, crackles, and bronchial sounds. Raw percentage of agreement (PA), Gwet's agreement coefficient type 1 (AC1), Krippendorff's alpha (K_a), and Fleiss kappa (K_{Fleiss}) were chosen as agreement indicators in the absence of a gold standard indicator to assess agreement. The different indicators were interpreted based on a priori reported benchmarks.

Results: The agreements were fair to good for almost all lung sound indicators. For the presence of any abnormal lung sound, the reported agreements (95% confidence intervals [CI]) were 0.781 (0.716–0.845), 0.646 (0.514–0.777), 0.403 (0.351–0.455), and 0.293 (0.137–0.493) for PA, AC1, K_a, and K_{Fleiss}, respectively. The same indicators were 0.769 (0.694–0.845), 0.615 (0.446–0.784), 0.426 (0.378–0.475), and 0.425 (0.293–0.563) for wheezes, 0.754 (0.685–0.823), 0.643 (0.503–0.782), 0.21 (0.146–0.275), and 0.208 (0.097–0.327) for crackles, and 0.636 (0.571–0.701), 0.345 (0.179–0.512), 0.182 (0.131–0.232), and 0.18 (0.081–0.279) for bronchial sound detections, respectively.

Conclusion and Clinical Importance: Agreement among raters auscultating calf respiratory sounds was higher than previously reported. However, improvement is still possible to increase auscultation agreement.

Abbreviations: AC1, Gwet agreement coefficient type 1; ACVIM, American College of Veterinary Internal Medicine; BCI, Bayesian credible interval; CI, confidence interval; ECBHM, European College of Bovine Health Management; IQR, interquartile range; K_a, Kripendorff alpha; K_{Fleiss}, Fleiss kappa; PA, percentage of agreement.

1 | Introduction

Lung auscultation is an integrative physical examination component for respiratory and non-respiratory conditions [1, 2]. Despite its wide availability, several criticisms have emerged based on various limitations including its subjectivity [3], especially when compared with other diagnostic procedures such as thoracic imaging [4–6]. In bovine medicine, the limitations of lung auscultation by veterinary practitioners have been described for auscultating young calves [7]. Inter-operator agreement for classifying normal versus abnormal lung auscultation was no better than chance among practitioners [7]. Recently, an attempt to overcome the limitations of thoracic auscultation by the standardization of lung sound nomenclature was reported [8]. Attempts also have been made to automate lung auscultation for feedlot calves by removing human subjective assessment [9, 10]. A particular challenge in bovine medicine is that the conditions in which auscultation is performed are challenging because of the presence of other animals or farm noises (e.g., ventilation fans, heating systems, movements, vocalizations of other animals). Currently, no consensus exists on which lung sounds may have acceptable agreement among veterinarians to allow their use in the detection of calves with bronchopneumonia.

Lung auscultation also has been defined as an art in human medicine because of the inherent difficulty in standardizing the definition of lung sounds [11]. Recently, several attempts have been made to improve lung sounds classification in human medicine [12]. These attempts have led to important simplifications of lung sound semiology and also have been used in a study on calf bronchopneumonia [8]. Moreover, lung sounds associated with better agreement among raters were identified [13, 14]. In a previous study in humans, a lung sound database was used to determine the agreement of auscultation findings by seven different groups of raters, including general practitioners from the Netherlands, Norway, Russia, and the United Kingdom, pulmonologists, researchers, and medical students [14]. Substantial heterogeneity of chance-corrected agreement was noted among groups of raters, with kappa varying between 0.27 and 0.97 for wheezes and between 0.2 and 0.58 for crackles. In companion animals, detection of adventitious sounds such as wheezes and crackles also has been investigated recently [15]. Kappa values for wheezes and crackles of 0.46 and 0.50 were obtained in dogs, whereas kappa values were 0.33 and 0.39 for the same sounds in cats. To our knowledge, a similar study has not been performed for lung sounds in cattle. Reliability and agreement are complex areas of medical research [16, 17]. Agreement is defined as how consistently the same indicator or measurement can be assessed repeatedly by different raters. Reliability helps determine if different patients can be distinguished from one another by the measured indicator (accounting for different sources of measurement errors) [18]. Agreement assessment remains a challenge because no universal agreement coefficient exists. Beyond raw percentage of agreement (PA), several chance-adjusted coefficients such as kappa, Krippendorff's alpha (K_a), and Gwet's agreement coefficients have been described for categorical indicators with advantages and disadvantages for each coefficient [19, 20].

Our main objective was to determine the agreement among veterinarians listening to different recorded lung sounds originating from field cases of lung auscultation. The main hypothesis was that some sounds can be detected with higher agreement than others. Knowing this crucial information could be helpful to tailor respiratory auscultation teaching and to identify the potential prognostic and diagnostic accuracy of sounds with good to very good inter-rater agreement in diagnosing bronchopneumonia. Secondary objectives were to determine if confidence in the diagnosis was associated with increased inter-rater agreement as well as the impact of each rater on the agreement values.

2 | Materials and Methods

2.1 | Lung Sounds Collection From Calves With Clinical Signs of Respiratory Disease

The research protocol was approved by the Ethical Committee of Research and Ethics of the Université de Montréal (24-Rech-2280). Lung sounds were collected using a digital stethoscope (Eko Littman Core, Eko Health, Emeryville, CA) in 50–80 kg milk-fed veal calves from November 2023 to March 2024. The four veal farms housed 495–815 animals. The collection took place during the first month upon arrival at the farms, a particularly high-risk period for respiratory disease outbreak [21]. Sound recording procedures were conducted while the calves were housed in individual pens before being grouped 4–5 weeks after arrival. The farms were visited between 9:00 a.m. and 10:00 a.m. for a maximum of 2 h total duration. This period was considered ideal because it occurred just after feeding (i.e., 7:00 a.m.–8:00 a.m.) when the calves were quiet and did not become excited by the expectation of being fed. The objective was to find and record various normal and abnormal lung sounds. The research team walked through the farm and observed calves that looked dull or had abnormal respiratory signs (e.g., dyspnea, cough, abnormal ear position, nasal discharge) to increase the probability of finding abnormal lung sounds. The selected calves were then rapidly auscultated by one author (S.B.) on both sides, and lung sounds generally were obtained from the third to the eighth intercostal space separating dorsal, mid, and ventral zones [8]. However, no systematic record of all sites was made because the objective was to focus on finding abnormal lung sounds. We did not a priori record the time taken per calf, but it was considered generally to be < 3 min. Calves with normal physical appearance also were included to collect normal lung sounds from the sixth to eighth intercostal space in the middle part of the right or left side. The lung sounds then were recorded using the Eko application via an iPhone 14 (Apple, Cupertino, CA) using a 15-s recording period. The research team tried to achieve the best conditions of auscultation for each sound under these circumstances.

2.2 | Lung Sounds Processing

The different recorded sounds then were extracted from the application portal (https://app.ekodevices.com) and converted into waveform audio format (.wav) files. Then, the audio files were imported using a free audio editor (Audacity, v 3.4.6; Audacity software; 1999–2014 Audacity Team: http://audacity.sourceforge.net/). The sound sequences then were edited to remove external noise (denoising) and amplified using a previously published method [22]. The audio samples then were stored on an external hard drive. The audio editor was used to create spectrograms of the recordings, and the Apowersoft

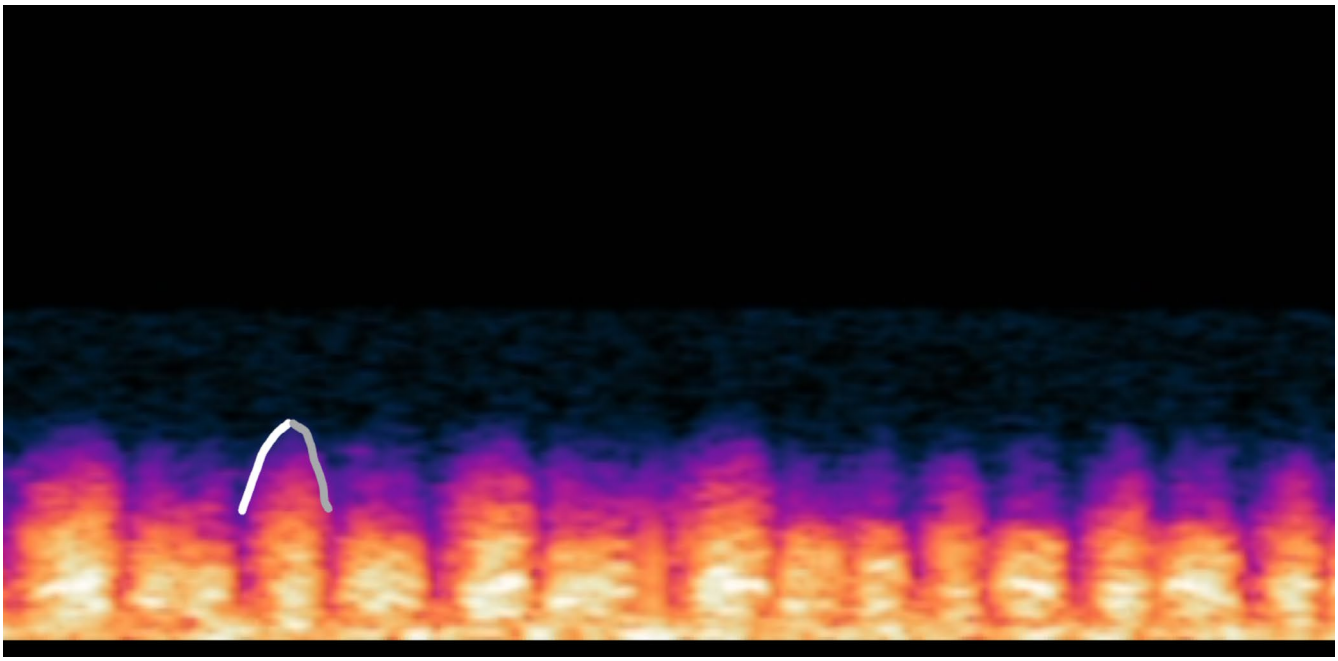

FIGURE 1 | Audio spectrogram of a calf with inspiratory and expiratory wheezes. The figure shows the onset of the inspiratory phase, characterized by an ascending profile (white line), whereas the expiratory phase immediately follows (gray line).

(Wangxu Technology Co. Ltd., Hong Kong) was used to extract the spectrogram videos from the audio editor (Figure 1). A moving timeline also was incorporated to link the sound with the corresponding sections of the spectrogram.

A short tutorial was built in an attempt to standardize training to assess lung sound auscultation. This tutorial (File S1) aimed to describe lung sounds using a pragmatic approach and was based on previously reported material for pulmonary auscultation of humans (Table 1). The tutorial was created using the Canva application, and the audio lecture was generated by Voicemaker (Yedap Technologies LLC, Sheridan, WY) and Vidnoz (Wise Reward Limited, Austin, TX) in French, English, and Portuguese using previously suggested nomenclature [12]. The tutorial was primarily focused on the definition of normal lung sounds, wheezes, crackles, and bronchial breath sounds. Other lung sounds such as pleural rub were not further investigated because of the low prevalence of detection of these sounds even in severe cases of pleuritis in sheep [23] and these sounds also appear quite infrequent in calves [8]. The tutorial was based on a review article [1] and a textbook [24]. The definitions used and associated findings are summarized in Table 1. The tutorial was 5 min in duration and practical, with examples of each type of lung sound provided and used as a baseline before participation in the project.

2.3 | Questionnaire Characteristics

A questionnaire was built to collect characteristics that could be associated with veterinarians recruited for lung sounds assessment, including gender, graduation year, percentage of time dedicated to cattle practice, as well as postgraduate diploma (clinically oriented [residency] vs. research-oriented [MSc or PhD]). For each lung sound loop, the veterinarian was asked about the presence of any abnormal lung sounds (absent vs. present vs. uncertain), for each type of abnormal lung sound heard (wheezes, crackles, bronchial sounds), and the evaluators' confidence in their choice using a 5-point Likert scale ranging from 1 (*not confident*) to 5 (*very confident*). The raters were free to indicate absence or presence of multiple abnormal sounds. A subjective gradation of the quality of lung sound recording also was scored by the veterinarians using a 5-point Likert scale from (1) *low quality* to (5) *excellent quality*. The online questionnaire (Google form, Alphabet, Mountain View, CA) was built by the first author (L.P.) and revised by the last author (S.B.) before being sent to the raters.

Ten different raters initially were selected for the study. These raters were chosen as a small sample that was representative of veterinarians with either an interest in individual cattle medicine or respiratory disease from different countries based on previous collaborations with the corresponding author (S.B.). Instructions were given, including how to listen to the tutorial (French, English, or Portuguese), which contained all information needed to complete the questionnaire.

2.4 | Sample Size and Statistical Analyses

2.4.1 | Sample Size Justification

Because of the limited information relative to veterinary auscultation, the sample size of the study was based on previous studies in human medicine reporting auscultation reliability based on 20 [14] to 30 different lung sounds records [25]. Additional sample size evaluation was based on simulation scenarios using designed R software [26] and the package power sample size. Several plausible scenarios then were built for kappa Fleiss (K_{Fleiss}) ranging from 0.5 to 0.9. Forty lung

TABLE 1 | Main lung sounds definitions and interpretation based on references [12] and [24].

Lung sound name	Synonym or old terms	Mechanism initiating the sound
Normal breathing sound	Vesicular murmur	Soft sound associated with airflow passage through the peripheral airways
Wheezes	Rhonchi	Continuous sound initiated by air passage throughout a narrowed bronchus
Adventitious sounds		
Crackles	Rales	Discontinuous sound associated either with air bubbling through secretions or a sudden rapid opening of the airways (also described as bubbling or popping sounds). Can be coarse or fine
Bronchial breath sounds		Increased inspiratory and expiratory sounds associated with transmission of large airway sound without attenuation to the area of auscultation due to the presence of consolidation around the large airways. Expiration is generally louder and longer than normal. Sound abnormal when heard over the lung area. Same intensity between inspiration and expiration
Pleural rub		Grating or squeaking sound associated with pleural inflammation between visceral and parietal pleura. Same intensity between inspiration and expiration

sounds loops were selected from a dataset where an approximately equal proportion of the four different types of lung sounds (normal, crackles, wheezes, bronchial sounds) were chosen by two authors not involved in auscultation scoring (L.P. and S.B.). These 2 authors selected 10 recorded loops from each category after consensus (wheezes, crackles, bronchial sounds, normal) based on the initial dataset of collected lung sounds. All of the sounds were randomly sorted before assessment by the raters.

2.4.2 | Statistical Analyses

All analyses were performed using the open access R software [26] using irrCAC and icr packages for agreement analysis [27, 28]. Descriptive statistics were added for rater characteristics as well as the quality score of sounds and confidence in the diagnosis.

Agreement was assessed using different agreement coefficients including PA, Gwet's agreement coefficient type 1 (AC1), K_a, and K_{Fleiss}. These indicators were chosen a priori in the absence of a gold standard measure of agreement. The PA was defined as acceptable if >0.75 using previously reported benchmarks [29, 30]. The AC1 is robust with respect to various Kappa bias paradoxes [20]. The AC1 and K_a can also be used in the presence of missing pairs or raters by contrast to K_{Fleiss}. The 95% confidence intervals (CIs) were obtained after normal approximation of the variance for PA and AC1 [19] or from 20 000 bootstrapped samples (K_a, K_{Fleiss}). The benchmarks used for chance adjusted agreement indicators were Altman a priori reported guidelines using poor (<0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), and very good (0.81–1.00) agreement depending on values of the agreement indices [31]. The respiratory sounds of interest were the presence of any abnormal sound and presence of wheezes, crackles, or bronchial sounds in the audio files assessed. A subsample analysis also was performed of lung sounds that the veterinarians rated with a confidence score of 4 or 5 on the Likert scale to determine the impact on agreement indices (without calculating K_{Fleiss} which cannot be used in the presence of missing pairs or raters caused by sounds selection). For each indicator, a sensitivity analysis also was performed for each indicator, removing the rater one-by-one to evaluate each rater's impact on the reported agreement indicators. The range of obtained values then was recorded.

A nonparametric Spearman ρ correlogram was used to assess inter-rater perception of lung sound quality and confidence in the auscultation diagnosis using "corrplot" and "reshape2" R packages [32, 33]. The interpretation of ρ values was based on a previous study [34] using negligible, weak, moderate, strong and very strong correlation for ρ values between 0 and 0.10, 0.10 and 0.39, 0.40 and 0.69, 0.70 and 0.89, and 0.90 and 1.00, respectively.

Rater-based prevalences of any abnormal sound, and for specific sounds such as wheezes, crackles, and bronchial sounds, were calculated. Differences among the proportions found by the different raters were assessed using a chi-squared test, using a statistically significant threshold at $p<0.05$.

3 | Results

Ten veterinarians participated in the sound assessment and scoring. Their median number of years since graduation was 19 (interquartile range [IQR], 16–25 years; range, 11–38 years). They graduated from Canadian ($n=3$), French ($n=3$), Italian ($n=3$), and Brazilian ($n=1$) veterinary schools. Their current area of practice was Québec, Eastern Canada ($n=5$), Italy ($n=2$), Sweden ($n=2$), and Brazil ($n=1$). Four of 10 raters were female veterinarians. All raters had a postgraduate diploma including clinical internship ($n=2$) or residency ($n=5$), board certification (American College of Veterinary Internal Medicine [ACVIM], $n=3$, European College of Bovine Health Management [ECBHM], $n=2$) or research qualification with a PhD ($n=6$).

The different indicators of agreement among raters for the different lung sound categories are summarized in Table 2 and in Figure 2.

TABLE 2 | Agreement indicators between 10 veterinarians scoring online 40 lung sound loops obtained from calves lung auscultation and subgroup analyses of sounds which were classified with high confidence by the raters.

Sound characteristics	Indicator	Estimate	95% CI	Sensitivity analysis[a]	Estimate	95% CI	Sensitivity analysis[a]
All loops					Sounds with confidence 4 or 5		
Any abnormal sound	PA	0.781	0.716–0.845	0.771–0.796	0.782	0.689–0.875	0.770–0.844
	AC1	0.646	0.514–0.777	0.629–0.678	0.649	0.466–0.831	0.627–0.761
	K_a	0.403	0.351–0.455	0.378–0.436	0.616	0.523–0.704	0.568–0.690
	K_{Fleiss}	0.293	0.137–0.493	0.273–0.322	—	—	—
Wheezes	PA	0.769	0.694–0.845	0.757–0.786	0.811	0.733–0.89	0.781–0.826
	AC1	0.615	0.446–0.784	0.604–0.639	0.681	0.511–0.852	0.627–0.705
	K_a	0.426	0.378–0.475	0.396–0.495	0.586	0.505–0.661	0.541–0.615
	K_{Fleiss}	0.425	0.293–0.563	0.394–0.493	—	—	—
Crackles	PA	0.754	0.685–0.823	0.744–0.771	0.801	0.722–0.880	0.770–0.843
	AC1	0.643	0.503–0.782	0.622–0.675	0.721	0.580–0.862	0.673–0.789
	K_a	0.21	0.146–0.275	0.178–0.257	0.315	0.191–0.433	0.275–0.335
	K_{Fleiss}	0.208	0.097–0.327	0.175–0.255	—	—	—
Bronchial sounds	PA	0.636	0.571–0.701	0.625–0.647	0.637	0.544–0.730	0.626–0.678
	AC1	0.345	0.179–0.512	0.316–0.387	0.308	0.098–0.519	0.286–0.370
	K_a	0.182	0.131–0.232	0.171–0.199	0.364	0.273–0.452	0.301–0.403
	K_{Fleiss}	0.18	0.081–0.279	0.161–0.197	—	—	—

Note: The color palette used depended on previously reported benchmarks for PA (acceptable if > 0.75, dark green; not acceptable ≤ 0.75, red). For other indicators the agreement was interpreted as using poor (< 0.20; red), fair (0.21–0.40; yellow), moderate (0.41–0.60; orange), good (0.61–0.80; pale green), or very good (0.81–1.00; dark green).

Abbreviations: AC1, Gwet agreement coefficient type 1; CI, confidence interval; K_a, Krippendorff's alpha; K_{Fleiss}, Fleiss kappa; PA, raw percentage of agreement.

[a]A sensitivity analysis was performed removing one-by-one each rater to see their individual impact on the agreement indicator. This range represented the obtained ranges from this sensitivity analysis.

Most of the calculated indicators were between fair and moderate for chance-adjusted agreement and good for raw agreement. The PA was acceptable for the presence of any abnormal sound, wheezes, or crackles but not for bronchial sounds. The estimates of agreement indicators remained stable in the sensitivity analysis obtained by removing one-by-one the results of each rater. This finding shows that the main results were relatively robust.

Within the same lung sound category, some variations were observed between the agreement indicators, with K_{Fleiss} and K_a generally lower than AC1. Focusing on lung sounds that were scored with high confidence by the raters, K_a increased for the presence of any abnormal sounds and detection of wheezes.

The correlations among the different raters of the perceived recording quality and confidence in the auscultation diagnosis are plotted in Figure 3. There were variations among raters regarding perceived recording quality and confidence in the diagnosis and the individual correlations among raters. The analysis of agreement focused on the sounds rated with high confidence by the raters (Table 2) did lead to important changes in agreement indicators. The median (IQR) ρ was 0.14 (0.05–0.30) for recording quality assessment and 0.15 (0–0.24) for confidence in the diagnosis, which was compatible with weak to moderate correlation in the majority of rater pairs. The proportions of particular abnormal sounds perceived by raters are indicated in Figure 4. No significant difference was observed between the individual proportion of wheezes ($p = 0.06$) or crackles ($p = 0.83$). However, significant heterogeneity was observed for bronchial sounds reporting ($p = 0.002$).

4 | Discussion

Our study provides insight into the agreement of veterinarians listening to the same lung sounds obtained from calves with various respiratory clinical signs recorded in a clinical setting. We found that, even using a short tutorial and specific conditions, the agreement among raters was fair to moderate, with most indicators lying between 0.2 and 0.8. These findings are in accordance with previous studies on lung auscultation in humans [14, 35, 36] and small animals [15], but much higher than what was observed in a previous study that found no agreement beyond chance comparing normal and abnormal respiratory auscultation without prior attempt to standardize lung sound terminology among 49 veterinarians auscultating among 8 to 10 calves [7].

Lung sounds terminology historically has been considered complex, which may explain why it has been described as an art with some subjectivity [37]. Terminology has been simplified in the

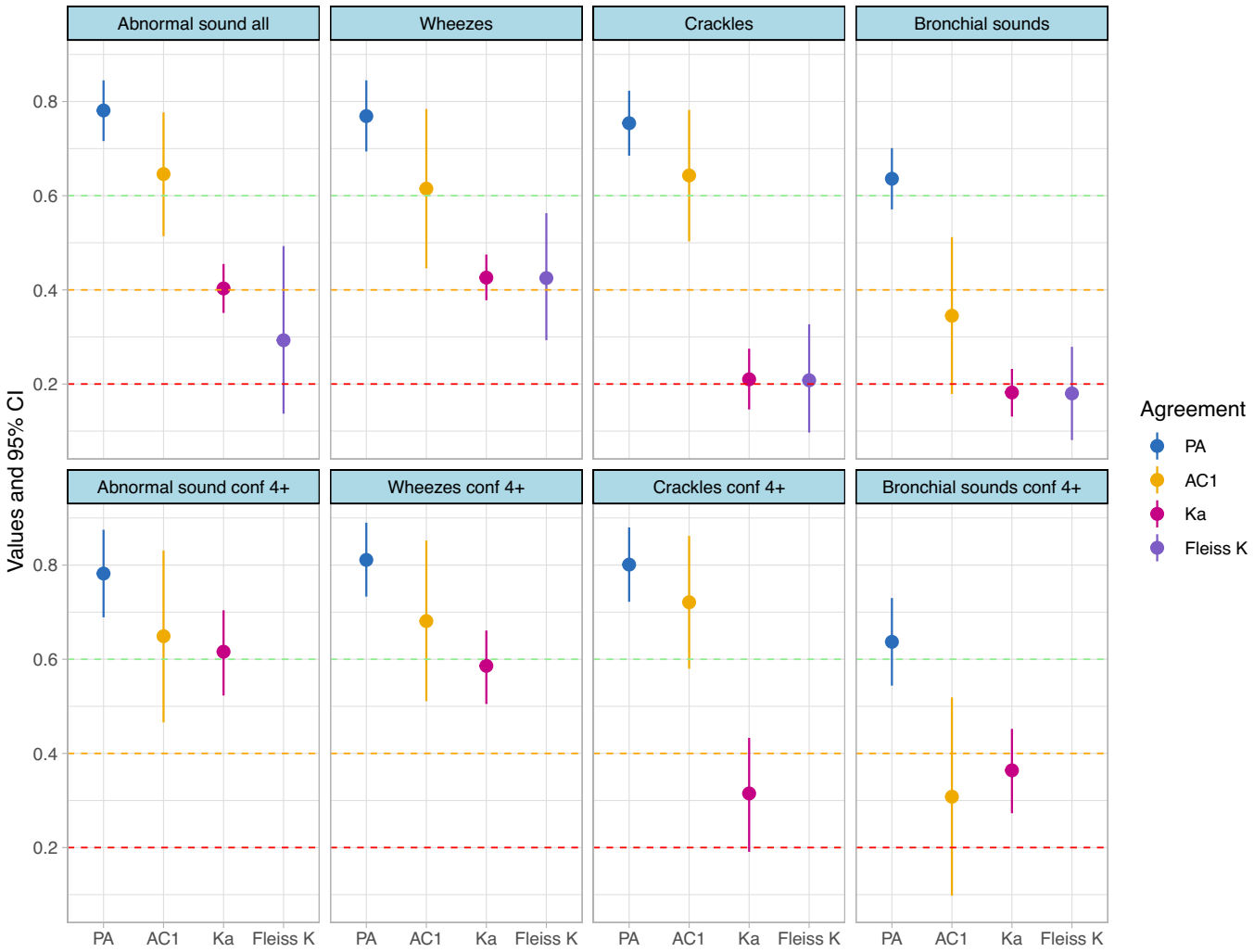

FIGURE 2 | Summary of different agreement indicators to detect wheezes, crackles, and bronchial sounds. The mean estimates and 95% confidence intervals are represented for raw percentage of agreement (PA), agreement coefficient type 1 (AC1), Krippendorff's alpha (K_a), and Fleiss kappa (K_{Fleiss}) for the 40 evaluated lung sound loops (upper panels) and focusing on loops that were classified with high confidence (Likert scale score 4 or 5 on 1–5 scale) by the veterinarian (lower panels). For this subset of analysis, Fleiss kappa could not be calculated due to missing pairs of raters' results.

previous 40 years in an attempt to increase agreement among different observers focusing on distinguishing between normal and abnormal sounds [1, 37]. We used a short online tutorial on lung sound nomenclature to standardize definitions of the different lung sounds under investigation. However, we did not record the total time that the raters spent using the tutorial or the time spent scoring all of the lung sounds, which could be covariates impacting reliability. Using a priori standardization of lung sound nomenclature previously has been shown to improve agreement among physical therapists auscultating lung sounds using a longer education session and discussion with specialists [3]. Using computer-assisted learning tools recently was proposed to avoid confusion in lung sound terminology [38]. Our tutorial was intended to be short so as to be compatible with what could be further developed for general practitioners who could be recruited to increase external validity. A structured module of auscultation learning adding some questions for self-assessment potentially would be relevant to increase the skills of veterinarians and veterinary students in performing lung auscultation. We focused on the presence of wheezes, crackles, and bronchial sounds, but did not include the presence of pleural rub and absence of respiratory sounds in our study because of the low frequency of these findings in dairy calves with bronchopneumonia [8]. In a previous study comparing lung ultrasonography to lung auscultation in sheep, pleural rub was not found in any of the six ewes with severe fibrinous pleuritis [23].

The different combinations of chance-corrected indicators gave acceptable agreement for the presence of any abnormal lung sound and wheezes. Lower agreement was observed for crackles, mostly for K_{Fleiss} and K_a. Crackles are discontinuous sounds (< 25 ms in duration) that can be observed with sudden changes of pressure (opening airways) or caused by air bubbles in secretions [24]. Fine and coarse crackles have been defined depending, among other factors, on the size of the affected airways as well as duration (5 vs. 15 ms) with shorter duration for fine crackles than for coarse crackles [39]. However, we did not a priori distinguish these two categories because of the inherent subjectivity of these definitions from a clinical point of view, as suggested by previous studies [14, 35] A distinction between fine and coarse crackles could have been made more precisely using spectrogram analyses and measurement of sound duration. However, we aimed to obtain inter-rater ratings that would be consistent with practical calf-side use of lung auscultation.

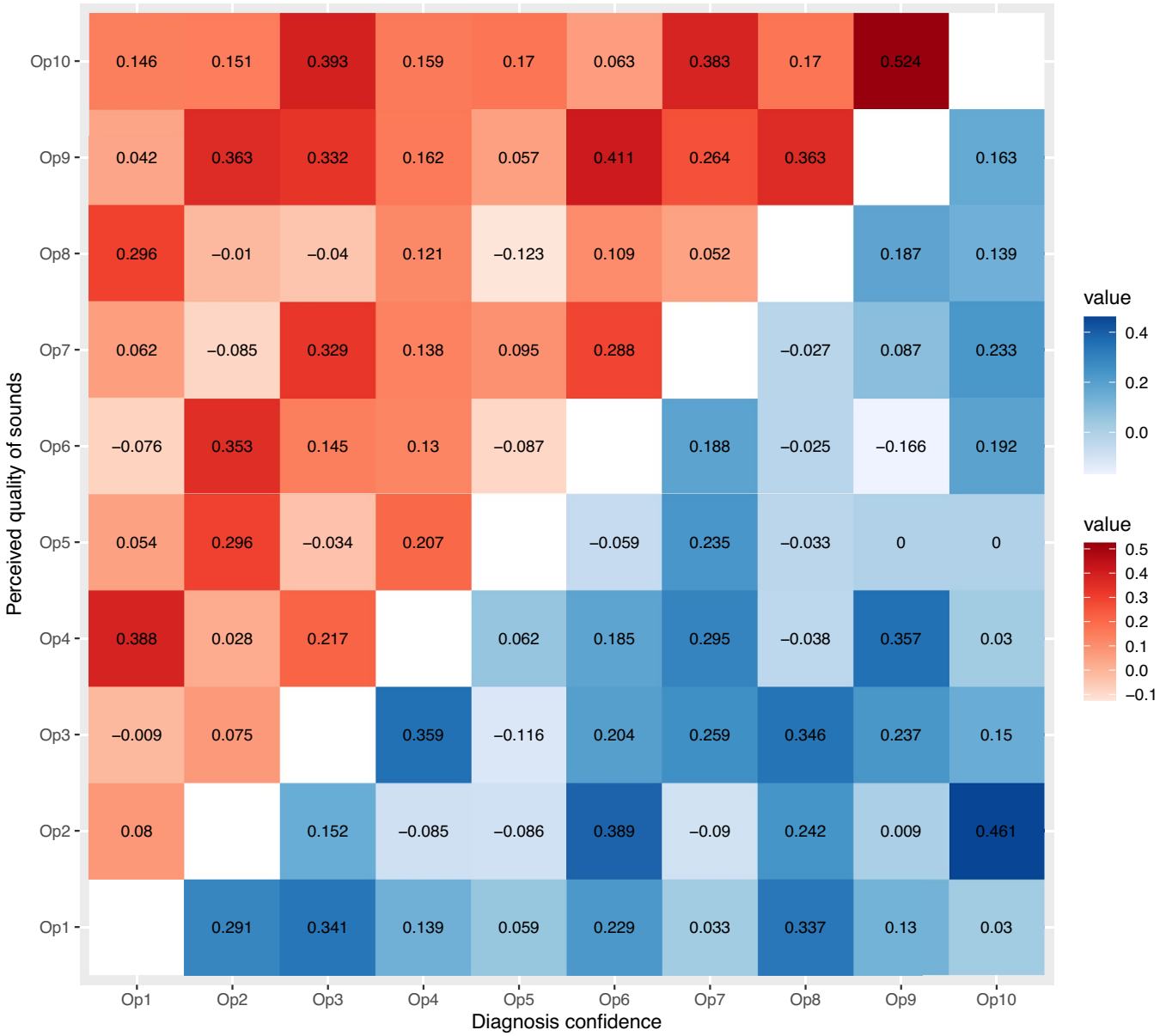

FIGURE 3 | Heatmap representing the perceived recording quality of sounds and confidence of diagnosis correlations (using Spearman ρ) between raters. The blue cells represent the individual correlations between the 10 operators' confidence (1–5 Likert scale) in their diagnosis of the 40 different loops (lower diagonal matrix). The red cells represent the individual correlations between the 10 operators' perception of the quality of the sounds (1–5 Likert scale).

In our opinion, simplifying the classification would broaden the use of lung auscultation.

Bronchial sounds had generally lower agreement than other lung sounds, even if selecting lung sounds that were classified with high confidence by the raters. Bronchial sounds typically are heard in the presence of lung consolidation when airflow within large bronchi is transmitted through airless consolidated lung tissue [39]. This pathological sound is detected on inspiration and expiration, with the same intensity (as compared to normal increased inspiratory sounds relative to expiratory sounds). These sounds were helpful in a previous study for diagnosing bronchopneumonia in calves. A high specificity (98.5%; 95% Bayesian credible interval [BCI]: 93.8%–99.9%) but lower sensitivity (67.9%; 95% BCI: 55.7–82.8) was observed for calves having at least one auscultation site with increased bronchial sounds or pleural friction rub, the latter anomaly being observed in only 1.2% of calves (4/330) [8]. A previous study analyzed answers from 187 physicians listening to recorded sounds from 24 different patients classified by experts as the gold standard. The percentage of correct detection of bronchial sounds ranged only from 15% to 30% by contrast to other sounds (crackles, 55%–75%; wheezes, 70%–90%), with higher accuracy for pneumologists than for pediatricians or medical students [11]. Inter-rater agreement for bronchial sounds detection was not different from chance only, in a study where K_{Fleiss} was only 0.034 for bronchial sounds (as compared with 0.704 for wheezes and 0.514 for crackles) based on the evaluation of 70 lung sounds by 7 specialist physicians [35]. Our study does not support the practical use of bronchial sounds as lung sounds with high agreement.

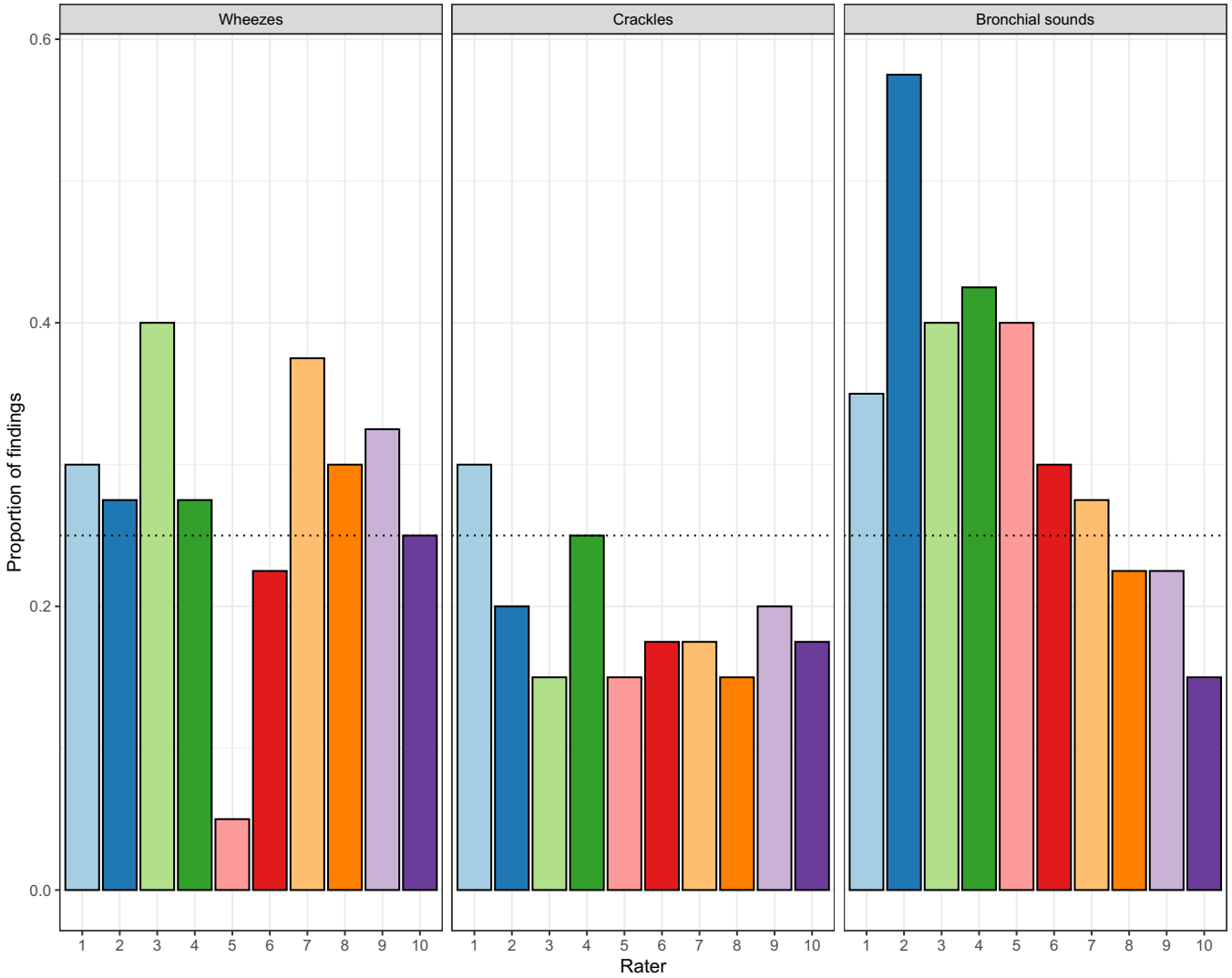

FIGURE 4 | Apparent prevalence of wheezes, crackles, and bronchial sounds of 10 different raters scoring 40 lung sound loops. The proportion of each abnormal sound finding by each rater is indicated. The dotted line represents the expected proportion of each lung sound in the dataset based on the initial selection of the sounds made by two authors not involved in agreement assessment.

The various indicators used for agreement determination gave heterogeneous results, which is not surprising. A commonly used benchmark for acceptability of agreement beyond chance is >0.6. However, the question of reliability still persists in the medical literature in the presence of multiple benchmark scales [40]. A small review of medical test agreement values concluded in the psychiatric field that, for clinical tests used in practice, observing a kappa above 0.8 would be almost miraculous; to see kappa between 0.6 and 0.8 would be a cause for celebration. A realistic goal is kappa between 0.4 and 0.6, while kappa between 0.2 and 0.4 would be acceptable [40]. This sentence emphasizes the importance of keeping these limitations in mind when reviewing medical tests used in daily practice. It is also important to use more than one agreement indicator because of the absence of a perfect indicator.

One strength of our study is that the audio samples submitted to the various raters lacked any information regarding the animal or the context of auscultation. This design enabled raters to separate their auscultatory findings from the calf's physical appearance, which could influence perceptions of lung sounds. For instance, examining a calf with an abnormal respiratory pattern might influence the veterinarian's assessment during auscultation. This scenario aligns with the concept of confirmation bias [41]. Our study design mitigated such bias. However, evaluating only a single sound might not accurately represent how a veterinarian conducts a thorough examination. Factors such as slightly repositioning the stethoscope to verify initial impressions, comparing right and left side sounds, and repeating the auscultation process may be crucial for a comprehensive assessment. In addition, the nature of recorded sounds presents inherent challenges in a posteriori evaluation of these recordings. Although appropriate settings were established for our study (e.g., analyzing sounds that always began with the inspiratory phase and the presence of a cursor in spectrogram sections), distinguishing between inspiration and expiration can be difficult, particularly in tachypneic calves. This issue may explain the limited agreement observed in the assessment of bronchial sounds, because an accurate distinction between inspiration and expiration is crucial for identifying this pathological sound. This distinction is more easily made when observing respiration in calves under natural conditions.

Thoracic auscultation has been a fundamental tool in veterinary medicine for diagnosing lower respiratory tract abnormalities. However, its limitations compared to modern imaging technologies have been well documented in human medicine [5, 6]. Similar limitations of thoracic auscultation also have been described in cattle, emphasizing the added value of lung ultrasonography over auscultation [42, 43]. A recent study showed more potential when using the same nomenclature than used in our study [8]. Our study sheds light on the challenges associated with thoracic auscultation, particularly concerning variability in agreement among raters. Evidence is currently lacking for the diagnostic and prognostic value of auscultation findings, except when using computer-assisted lung auscultation, which can provide an accurate diagnosis [9] and prognosis [10] in feedlot calves using a different algorithm from classical lung auscultation nomenclature. Moreover, the potential of digital lung auscultation as a promising tool for enhancing the accuracy and agreement of this examination has been emphasized [44]. However, evidence regarding its use in clinical settings remains limited, as noted in a recent systematic review on childhood pneumonia [45].

We were interested in investigating if agreement increased with the confidence the raters had in their diagnosis. However, because of the limited correlation between confidence among the raters, this hypothesis could not be confirmed. We tried to calculate agreement indicators robust to missing data (PA, AC1, K_a) in the subset of sounds that were scored with high confidence by the raters, but did not observe any relevant effect on the value of these indicators.

Our study had several limitations. First, despite our efforts to standardize lung sound terminology using a short tutorial, it remains uncertain whether this tutorial will effectively change the preconceived notions of a diversified group of experienced clinicians. We did not have a control group without access to the tutorial. The raters did not have a recent audiogram to record eventual hearing losses of some sounds range, which also may be a potential cause of disagreement. Differences in sounds definition and interpretation also could be associated with the native language due to either definition of sounds [46] or use of the audio spectrum in the different languages [47]. Informal feedback from the raters indicated some difficulty in separating the sounds from the everyday context of the routine clinical examination. In particular, this difficulty was attributed to the inability to visualize the calves during assessment of the recorded sounds. Because of the limited number of raters involved in the study, we could not evaluate rater characteristics associated with agreement indicators. This evaluation would be another potentially important future study to perform. However, our sensitivity analysis removing raters one-by-one to recalculate the agreement indicators did not identify important changes, showing that our findings were not affected by one particular rater. Our study did not enable us to determine if we would be able to teach future veterinarians with no prior education on lung sound terminology a simplified lung sound terminology using the same tutorial to enhance their agreement and increase their confidence in assessing lung sounds in cattle.

In conclusion, our findings indicated that it is possible to get fair to good agreement for the detection of any abnormal sound, wheezes, and crackles, whereas agreement was poor to fair for the detection of bronchial sounds. The potential of using a recorded dataset of lung sounds to standardize lung sound classification and improve clinical teaching in bovine medicine should be investigated more thoroughly.

Disclosure

Authors declare no off-label use of antimicrobials.

Ethics Statement

Approval by the ethical committee of research and ethics of the Université de Montréal (24-Rech-2280) to protocol to obtain the lung sounds from the auscultated calves. Authors declare human ethics approval was not needed.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of Lung Auscultation," *New England Journal of Medicine* 370, no. 8 (2014): 744–751, https://doi.org/10.1056/NEJMra1302901.

2. T. O. Cheng, "How Laënnec Invented the Stethoscope," *International Journal of Cardiology* 118, no. 3 (2007): 281–285, https://doi.org/10.1016/j.ijcard.2006.06.067.

3. D. Brooks and J. Thomas, "Interrater Reliability of Auscultation of Breath Sounds Among Physical Therapists," *Physical Therapy* 75, no. 10 (1995): 1082–1088, https://doi.org/10.1093/ptj/75.12.1082.

4. R. L. Murphy, "In Defense of the Stethoscope," *Respiratory Care* 53, no. 3 (2008): 355–369.

5. J. Lovrenski, S. Petrović, S. Balj-Barbir, R. Jokić, and G. Vilotijević-Dautović, "Stethoscope vs. Ultrasound Probe—Which Is More Reliable in Children With Suspected Pneumonia?," *Acta Medica Academica* 45, no. 1 (2016): 39–50, https://doi.org/10.5644/ama2006-124.155.

6. E. G. M. Cox, G. Koster, A. Baron, et al., "Should the Ultrasound Probe Replace Your Stethoscope? A SICS-I Sub-Study Comparing Lung Ultrasound and Pulmonary Auscultation in the Critically Ill," *Critical Care* 24, no. 1 (2020): 14, https://doi.org/10.1186/s13054-019-2719-8.

7. B. Pardon, S. Buczinski, and P. R. Deprez, "Accuracy and Inter-Rater Reliability of Lung Auscultation by Bovine Practitioners When Compared With Ultrasonographic Findings," *Veterinary Record* 185, no. 4 (2019): 109, https://doi.org/10.1136/vr.105238.

8. A. Boccardo, S. Ferraro, G. Sala, V. Ferrulli, D. Pravettoni, and S. Buczinski, "Bayesian Evaluation of the Accuracy of a Thoracic Auscultation Scoring System in Dairy Calves With Bronchopneumonia Using a Standard Lung Sound Nomenclature," *Journal of Veterinary Internal Medicine* 37, no. 4 (2023): 1603–1613, https://doi.org/10.1111/jvim.16798.

9. A. V. Mang, S. Buczinski, C. W. Booker, and E. Timsit, "Evaluation of a Computer-Aided Lung Auscultation System for Diagnosis of Bovine Respiratory Disease in Feedlot Cattle," *Journal of Veterinary Internal Medicine* 29, no. 4 (2015): 1112–1116, https://doi.org/10.1111/jvim.12657.

10. C. W. Booker, G. K. Jim, T. M. Grimson, B. K. Widman, and J. N. Nickell, "Association Between Computer-Aided Lung Auscultation and Treatment Failure Risk in Calves Treated for Respiratory Disease," *Canadian Veterinary Journal* 62, no. 5 (2021): 511–514.

11. H. Hafke-Dys, A. Bręborowicz, P. Kleka, J. Kocinski, and A. Biniakowski, "The Accuracy of Lung Auscultation in the Practice of

Physicians and Medical Students," *PLoS One* 14, no. 8 (2019): e0220606, https://doi.org/10.1371/journal.pone.0220606.

12. H. Pasterkamp, P. L. Brand, M. Everard, L. Garcia-Marcos, H. Melbye, and K. N. Priftis, "Towards the Standardisation of Lung Sound Nomenclature," *European Respiratory Journal* 47, no. 3 (2016): 724–732, https://doi.org/10.1183/13993003.01132-2015.

13. H. Melbye, L. Garcia-Marcos, P. Brand, M. Everard, K. Priftis, and H. Pasterkamp, "Wheezes, Crackles and Rhonchi: Simplifying Description of Lung Sounds Increases the Agreement on Their Classification: A Study of 12 Physicians' Classification of Lung Sounds From Video Recordings," *BMJ Open Respiratory Research* 28, no. 1 (2016): e000136, https://doi.org/10.1136/bmjresp-2016-000136.

14. J. C. Aviles-Solis, S. Vanbelle, P. A. Halvorsen, et al., "International Perception of Lung Sounds: A Comparison of Classification Across Some European Borders," *BMJ Open Respiratory Research* 4, no. 1 (2017): e000250, https://doi.org/10.1136/bmjresp-2017-000250.

15. M. Domínguez-Ruiz, C. R. Reinero, A. Vientos-Plotts, M. E. Grobman, D. Silverstein, and K. Le Boedec, "Interclinician Agreement on the Recognition of Selected Respiratory Clinical Signs in Dogs and Cats With Abnormal Breathing Patterns," *Veterinary Journal* 277 (2021): 105760, https://doi.org/10.1016/j.tvjl.2021.105760.

16. R. Hernaez, "Reliability and Agreement Studies: A Guide for Clinical Investigators," *Gut* 64, no. 7 (2015): 1018–1027, https://doi.org/10.1136/gutjnl-2014-308619.

17. H. C. W. de Vet, C. B. Terwee, L. B. Mokkink, and D. L. Knol, *Measurement in Medicine: A Practical Guide* (Cambridge University Press, 2011).

18. H. C. W. de Vet, C. B. Terwee, D. L. Knol, and L. M. Bouter, "When to Use Agreement Versus Reliability Measures," *Journal of Clinical Epidemiology* 59, no. 10 (2006): 1033–1039, https://doi.org/10.1016/j.jclinepi.2005.10.015.

19. K. L. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (Advanced Analytics LLC, 2014).

20. H. C. de Vet, L. B. Mokkink, C. B. Terwee, O. S. Hoekstra, and D. L. Knol, "Clinicians Are Right Not to Like Cohen's Kappa," *BMJ* 346 (2013): f2125, https://doi.org/10.1136/bmj.f2125.

21. B. Pardon, K. de Bleecker, M. Hostens, J. Callens, J. Dewulf, and P. Deprez, "Longitudinal Study on Morbidity and Mortality in White Veal Calves in Belgium," *BMC Veterinary Research* 8 (2012): 26, https://doi.org/10.1186/1746-6148-8-26.

22. E. D. McCollum, D. E. Park, N. L. Watson, et al., "Listening Panel Agreement and Characteristics of Lung Sounds Digitally Recorded From Children Aged 1–59 Months Enrolled in the Pneumonia Etiology Research for Child Health (PERCH) Case–Control Study," *BMJ Open Respiratory Research* 4, no. 1 (2017): e000193, https://doi.org/10.1136/bmjresp-2017-000193.

23. P. Scott, D. Collie, B. McGorum, and N. Sargison, "Relationship Between Thoracic Auscultation and Lung Pathology Detected by Ultrasonography in Sheep," *Veterinary Journal* 186, no. 1 (2010): 53–57, https://doi.org/10.1016/j.tvjl.2009.07.020.

24. J. S. Coviello, *Auscultation Skills: Breath & Heart Sounds* (Lippincott Williams & Wilkins, 2013).

25. J. C. Aviles-Solis, I. Storvoll, S. Vanbelle, and H. Melbye, "The Use of Spectrograms Improves the Classification of Wheezes and Crackles in an Educational Setting," *Scientific Reports* 10, no. 1 (2020): 8461, https://doi.org/10.1038/s41598-020-65354-w.

26. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).

27. K. L. Gwet and M. K. L. Gwet, "Package "irrCAC"," in *Computing Chance-Corrected Agreement Coefficients (CAC)* (CRAN, 2019).

28. A. Staudt, P. L'Ecuyer, and C.-H. Chan, "Package icr," CRAN, 2024.

29. C. C. Burn and A. A. Weir, "Using Prevalence Indices to Aid Interpretation and Comparison of Agreement Ratings Between Two or More Observers," *Veterinary Journal* 188, no. 2 (2011): 166–170, https://doi.org/10.1016/j.tvjl.2010.04.021.

30. S. Buczinski, C. Buathier, A. M. Belanger, H. Michaux, N. Tison, and E. Timsit, "Inter-Rater Agreement and Reliability of Thoracic Ultrasonographic Findings in Feedlot Calves, With or Without Naturally Occurring Bronchopneumonia," *Journal of Veterinary Internal Medicine* 32, no. 5 (2018): 1787–1792, https://doi.org/10.1111/jvim.15257.

31. D. G. Altman, *Practical Statistics for Medical Research* (Chapman and Hall, 1991).

32. T. Wei, V. Simko, M. Levy, Y. Xie, Y. Jin, and J. Zemla, "Package 'Corrplot'," *Stat* 56 (2017): e24.

33. H. Wickham, "Reshaping Data With the Reshape Package," *Journal of Statistical Software* 21, no. 12 (2007): 1–20, https://doi.org/10.18637/jss.v021.i12.

34. P. Schober, C. Boer, and L. A. J. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesthesia and Analgesia* 126, no. 5 (2018): 1763–1768, https://doi.org/10.1213/ANE.0000000000002864.

35. D. Magor, E. Berkov, D. Siomin, et al., "Interpretation of Heart and Lungs Sounds Acquired via Remote, Digital Auscultation Reached Fair-to-Substantial Levels of Consensus Among Specialist Physicians," *Diagnostics* 13, no. 19 (2023): 3153, https://doi.org/10.3390/diagnostics13193153.

36. J. Benbassat and R. Baumal, "Narrative Review: Should Teaching of the Respiratory Physical Examination Be Restricted Only to Signs With Proven Reliability and Validity?," *Journal of General Internal Medicine* 25, no. 8 (2010): 865–872, https://doi.org/10.1007/s11606-010-1327-8.

37. R. Curtis, L. Viel, S. McGuirk, O. M. Radostits, and F. W. Harris, "Lung Sounds in Cattle, Horses, Sheep and Goats," *Canadian Veterinary Journal* 27, no. 4 (1986): 170–172.

38. A. Bohadana, H. Azulai, A. Jarjoui, G. Kalak, and G. Izbicki, "Influence of Observer Preferences and Auscultatory Skill on the Choice of Terms to Describe Lung Sounds: A Survey of Staff Physicians, Residents and Medical Students," *BMJ Open Respiratory Research* 7, no. 1 (2020): e000564, https://doi.org/10.1136/bmjresp-2020-000564.

39. M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the Respiratory System," *Annals of Thoracic Medicine* 10, no. 3 (2015): 158–168, https://doi.org/10.4103/1817-1737.160831.

40. H. C. Kraemer, D. J. Kupfer, D. E. Clarke, W. E. Narrow, and D. A. Regier, "DSM-5: How Reliable Is Reliable Enough?," *American Journal of Psychiatry* 169, no. 1 (2012): 13–15.

41. K. van den Berge and S. Mamede, "Cognitive Diagnostic Error in Internal Medicine," *European Journal of Internal Medicine* 24, no. 6 (2013): 525–529, https://doi.org/10.1016/j.ejim.2013.03.006.

42. S. Buczinski, G. Forté, D. Francoz, and A. M. Bélanger, "Comparison of Thoracic Auscultation, Clinical Score, and Ultrasonography as Indicators of Bovine Respiratory Disease in Preweaned Dairy Calves," *Journal of Veterinary Internal Medicine* 28, no. 1 (2014): 234–242, https://doi.org/10.1111/jvim.12251.

43. S. Buczinski, J. Ménard, and E. Timsit, "Incremental Value (Bayesian Framework) of Thoracic Ultrasonography Over Thoracic Auscultation for Diagnosis of Bronchopneumonia in Preweaned Dairy Calves," *Journal of Veterinary Internal Medicine* 30, no. 4 (2016): 1396–1401, https://doi.org/10.1111/jvim.14361.

44. D. Bardou, K. Zhang, and S. M. Ahmad, "Lung Sounds Classification Using Convolutional Neural Networks," *Artificial Intelligence in Medicine* 88 (2018): 58–69, https://doi.org/10.1016/j.artmed.2018.04.008.

45. S. Ahmed, S. Sultana, A. M. Khan, et al., "Digital Auscultation as a Diagnostic Aid to Detect Childhood Pneumonia: A Systematic Review," *Journal of Global Health* 12 (2022): 04033, https://doi.org/10.7189/jogh.12.04033.

46. I. Fernandez-Prieto, C. Spence, F. Pons, and J. Navarra, "Does Language Influence the Vertical Representation of Auditory Pitch and Loudness?," *i-Perception* 8, no. 3 (2017): 2041669517716183, https://doi.org/10.1177/2041669517716183.

47. K. Sekiyama, "Influence of Language Backgrounds on Audiovisual Speech Perception Across the Lifespan," *Acoustical Science and Technology* 41, no. 1 (2020): 37–38, https://doi.org/10.1250/ast.41.37.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **File S1.** Presentation about lung auscultation. This video provides a step-by-step walkthrough of the procedures.