Evaluation of inter-rater agreement of the clinical signs used to diagnose bovine respiratory disease in individually housed veal calves

J. Berman,1* D. Francoz,1 A. Abdallah,1,2 S. Dufour,3 and S. Buczinski1

1Département des sciences cliniques, Faculté de médecine vétérinaire, Université de Montréal, Saint-Hyacinthe, QC J2S 2M2, Canada
2Department of Animal Medicine, Faculty of Veterinary Medicine, Zagazig University, Zagazig, City 44511, Egypt
3Département de Pathologie et Microbiologie, Faculté de médecine vétérinaire, Université de Montréal, St-Hyacinthe, QC J2S 2M2, Canada

ABSTRACT

In dairy calves raised for veal, typical clinical signs of bovine respiratory disease (BRD) are ocular discharge, nasal discharge, ear droop or head tilt, abnormal respiration, cough, and increased rectal temperature. Despite the existence of several clinical scoring systems, there are few studies on the variability of human recognition of individual BRD clinical signs. The objective of this study was therefore to assess the inter-rater agreement of BRD clinical signs in veal calves. We hypothesized that BRD clinical signs were not detected equally between veterinarians, technicians, and producers of the veal industry and that some clinical signs have higher inter-rater agreement than others. During 2017–2018, we prospectively recorded 524 videos of physical examinations of random veal calves from 48 different batches in Québec, Canada. A researcher, not involved in the inter-rater assessment, classified each video as presence/absence of each BRD clinical sign except rectal temperature. For each of the 5 clinical signs, 15 videos with and 15 videos without the clinical signs were randomly selected to avoid kappa paradoxes. Those 30 videos were then presented in a random order to experienced raters of BRD in veal calves: 6 veterinarians, 6 technicians, and 6 producers. The raters assessed the clinical signs using scores based on the Wisconsin and California scoring system with modifications (0 = absent, 1 = mild, 2 = moderate, 3 = severe for nasal discharge, ocular discharge, and ear droop or head tilt; and 0 = absent, 1 = moderate, 2 = severe for abnormal respiration and induced cough). We used median percentage agreement (Pa), median Cohen’s kappa (κ), and Gwet’s agreement coefficient 1 (AC1) to assess inter-rater agreement. The effect of scale combination was also tested to determine the optimal combination (4-scale 0/1/2/3 vs. 3-scale 0/1/2 vs. 2-scale 0/1,2,3; 0,1/2,3; or 0/1,2). The differences of inter-rater agreement between veterinarians, technicians, and producers were estimated by a Wilcoxon rank-sum test. The 2-scale combination (0,1/2,3 or 0/1,2) had the highest inter-rater agreement for all clinical signs. With this combination, induced cough was the clinical sign with the highest inter-rater agreement (Pa = 0.93; κ = 0.79; AC1 = 0.87) and abnormal respiration was the sign with the lowest inter-rater agreement (Pa = 0.77; κ = 0.20; AC1 = 0.74). According to Pa and AC1 values, the 2-scale inter-rater agreement of the 5 clinical signs was good (value > 0.6). According to κ, only ear droop or head tilt and induced cough had a substantial 2-scale inter-rater agreement (κ > 0.6). In general, the 2-scale inter-rater agreement was better among veterinarians than among technicians and producers, except for the ear droop/head tilt, where agreement was better among producers. We concluded that with severity scores assessed on a scale of 2 (0,1/2,3 or 0/1,2), the inter-rater agreement of BRD clinical signs was variable according to the sign in veal calves. BRD clinical signs were not detected equally between veterinarians, technicians, and producers of the veal industry. Future research could determine if this discrepancy could be improved by standardization training.

Key words: reliability, interoperator, pneumonia, kappa, cattle

INTRODUCTION

Veal production is a major outlet for dairy or cross-breed calves from the dairy industry (Sans and Font-guyon, 2009). Bovine respiratory disease (BRD) is a multifactorial infectious disease of the upper and lower respiratory tract (Woolums, 2015) highly prevalent in this industry (van der Mei and van den Ingh, 1987; Leruste et al., 2012; Pardon et al., 2012b). Economically, BRD in veal calves is responsible for mortality (Sargeant et al., 1994a,b; Pardon et al., 2013; Lava et al., 2016a), lighter carcass weight, and lower carcass...
quality at slaughter (van der Mei and van den Ingh, 1987; Pardon et al., 2013). Importantly, BRD is also a major cause of antimicrobial use in veal calves (Pardon et al., 2012a,b; Pardon et al., 2015). An accurate and reliable BRD individual and group-level diagnosis is therefore essential to reduce economic losses and use more judiciously antimicrobials.

Currently, BRD diagnosis and treatment decisions in the veal industry are generally based on examination for BRD clinical signs, performed mostly by producers on a daily basis, along with supervision and occasional intervention of veterinarians and technicians (i.e., veal calf experts that visits the farm each week to perform a quality inspection of the animals; Leruste et al., 2012; Pardon et al., 2012a; Lava et al., 2016b). However, disagreement currently exists between BRD clinical signs detection during the fattening period and lung lesion prevalence at slaughter in veal calves (5.0% of batches with BRD clinical signs vs. 51.7% of batches with lesions at slaughter, respectively; Leruste et al., 2012). If a BRD clinical sign has a poor inter-rater agreement, it may result in inaccurate diagnoses across examiners and practice settings (Kottner et al., 2011). Moreover, previous works in dairy cattle have shown differences of clinical examination findings between veterinarians and producers (Virtala et al., 1999; Olson et al., 2019). It is therefore essential to define reliable BRD clinical signs for producers, technicians and veterinarians working in the veal industry.

Currently, common specific BRD clinical signs are ocular discharge, nasal discharge, ear droop or head tilt, cough, abnormal respiration, and hyperthermia (Lago et al., 2006; McGuirk, 2008; Love et al., 2014). To date, previous studies reporting inter-rater agreement for BRD detection showed slight to fair agreement (Amrine et al., 2013; Buczinski et al., 2016). Robust individual BRD clinical signs inter-rater agreement estimations are needed to define which BRD clinical signs are reliable enough among veterinarians, producers, and technicians to be incorporated in a scoring system and provide clear guidance for a practical examination process to standardize case definition.

The main objective of this study was, therefore, to evaluate the inter-rater agreement of clinical signs used to diagnose BRD in individually housed veal calves. Our hypothesis was that the different BRD clinical signs (ocular discharge, nasal discharge, ear droop or head tilt, abnormal respiration, induced cough) are not detected equally between different examiners (i.e., veterinarians, technicians, and producers) working in veal farming, but that some clinical signs are robust enough (i.e., have high inter-rater agreement) to be included in a clinical scoring system for BRD monitoring.

**MATERIALS AND METHODS**

To estimate the inter-rater agreement of ocular discharge, nasal discharge, ear droop or head tilt, abnormal respiration, and spontaneous or induced cough, we designed a simulation study where videos of those BRD clinical signs were assessed by producers, technicians, and veterinarians. This project was designed according to the guidelines for reporting reliability and agreement studies (Kottner et al., 2011; Supplemental Table S1, https://figshare.com/s/53e9c054179876857812). This study was approved by the Comité d’Éthique de l’Université de Montréal (17_rech-1898).

**Collection of the Videos**

**Selection of Calves.** During a cross-sectional study performed from October 1, 2017, to December 20, 2018, 524 calves were prospectively and randomly recruited from 48 batches of 36 commercial rosé or white veal calf fattening units in Québec (Délinax, Veaux Lourds L’té). Single batch of calves (group of calves that arrived at the farm at the same time to be fattened and slaughtered together) was considered for each farm (Leruste et al., 2012), and the selected batches were distributed across all 4 seasons. The veal calves were mostly obtained from multiple local dairy farms through auction markets in the first 2 weeks of life. The calves were housed individually during the early feeding period (up to 2 mo). Because the number of calves present in every unit was known before the visit and each individual box was numbered on the farm, 10 individual veal calves were randomly chosen [RANDOM function in Excel 2016 (Microsoft Corp.)]. Whenever a box was empty due to early mortality, the next random number was selected. During the visit, if a nonrandomly selected calf had potentially interesting relevant clinical signs on the farm, this additional calf was also filmed to potentially adjust for prevalence of normal or abnormal BRD clinical signs and to avoid prevalence bias impact on agreement measures (Byrt et al., 1993).

**Recording of Videos.** Data were collected 2 weeks after the calves’ arrival at the fattening unit to intervene during the early main clinical BRD peak incidence (Miller et al., 1980; Pardon et al., 2012b). At that time, calves were housed in individual duckboard pens (baby boxes). Data were collected in the morning (between 0900 and 1100 h), between meals. Each calf was filmed with a portable camera (GoPro Hero 5) for a period of 2 to 3 min to assess the BRD clinical signs previously reported in the literature, such as nasal discharge, ocular discharge, ear droop or head tilt, abnormal res-
piration, and spontaneous or induced cough (McGuirk, 2008; Love et al., 2014). Precisely, nasal discharge, ocular discharge, and ear droop or head tilt were recorded by filming the head of each calf when the calf was alone in its box. Abnormal respiration was recorded by focusing on the border between the abdominal and thoracic walls of each calf. To maintain a short recording period similar to that of a day-to-day screening examination process, spontaneous cough could not be consistently recorded. An operator (A.A. or S.B.) entered the baby box and induced cough was recorded by filming the calf’s response to manual compression of the trachea. Videos had a resolution of 1,920 × 1,080 pixels at a frame rate of 60 frames per setting.

Selection of Videos. Figure 1 shows the selection steps of the videos. First, a single operator (J.B.), not included as a future rater, classified each video in the following categories in step 1 (Figure 1): normal (with no abnormal BRD clinical signs), presence of nasal discharge, presence of ocular discharge, presence of ear droop or head tilt, presence of abnormal respiration, or presence of induced cough. One video could be in several categories.

In step 2, 100 videos were randomly selected in each sign’s category using the RANDOM function in Excel (Microsoft Corp; Figure 1). If the number of videos was less than 100 in one category, all the videos were kept for the next step.

Each selected video was then shortened with movie software (Windows Movie Maker, Windows 8, Microsoft Corp.) to isolate the targeted BRD clinical sign of the category. The 100 videos in the normal category were similarly shortened for each clinical sign and divided into groups of 20 control videos (supposedly negative based on initial screening) to assess the presence of nasal discharge, 20 control videos to assess the presence of ocular discharge, 20 control videos to assess the presence of ear droop/head tilt, 20 control videos to assess the presence of abnormal respiration, and 20 control videos to assess the presence of induced cough. Sound was preserved for all videos, and final videos for a specific clinical sign were 10 to 30 s in duration.

Each video was scored by the same experienced operator (J.B.), who was not included as a future rater, using the score presented in Table 1. A video was included if the scoring of the target BRD clinical sign was judged possible by the operator (J.B.; e.g., visualization of 2 eyes for the scoring of ocular discharge or 2 nostrils for the scoring of nasal discharge; step 3, Figure 1). Moreover, videos with excessive darkness, contrast problems or excessive movement of the calf were not included in this study because it was anticipated that it could

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Flowchart of the selection (4 steps) of the videos from the 524 filmed veal calves per category of bovine respiratory disease (BRD) clinical signs: nasal discharge, ocular discharge, ear droop or head tilt, abnormal respiration, and induced cough.
A video was also excluded if the video was already selected in other category, and if external elements could influence the assessment of the BRD clinical signs (written indication of treatment on the box, other apparent clinical signs impossible to dissociate with the target BRD clinical sign on the video (e.g., presence of nasal discharge in the ocular discharge category) or oral indication (i.e., examiners talking about a particular clinical sign) during the physical examination).

Assessment of the Videos by the Raters

Videos were randomly selected within the database for a total of 30 videos per clinical sign (step 4, Figure 1). Because inter-rater agreement estimated by Cohen’s kappa ($\kappa$) can be affected when one of the scoring categories being rated has low prevalence (so-called $\kappa$ paradox), the 2-by-2 table margins were initially set to have a prevalence of 0.5 as initially classified [i.e., each set of 30 videos per BRD clinical sign included approximately 15 videos initially classified by J.B. without and 15 videos classified with the clinical sign (score 1, 2 or 3)].

Selection and Training of the Raters. A convenience sample was made up of 6 veterinarians, 6 technicians, and 6 producers. All the raters had various practical and relevant experience in the assessment of BRD in veal calves (Table 2). Before presenting the videos, each rater received basic training (slide show presentation) with specific explanations of how to score the severity of each BRD clinical sign (Table 1). Briefly, the assessment of nasal discharge, ocular discharge and ear droop or head tilt was performed using the 4-level scale of the Wisconsin scoring system chart (WI) as published by McGuirk (2008). The cough assessment was modified from the WI score because of the difficulty recording grade 3 (spontaneous cough) in this study using a relatively short duration of observation. The California score (CA; Love et al., 2014) was modified for assessing abnormal respiration consisting of a score of 0 if the respiratory pattern was judged normal, of 1 if tachypnea without dyspnea was present, and of 2 if dyspnea was present.

Assessment of the Videos. Each video was then individually scored by each recruited rater via the tutorial software Edpuzzle (Mischel, 2019). The videos were available online, and each rater could individually visualize them anytime and anywhere via his or her computer. The assessment of videos was consequently performed independently and blindly of the other raters’ responses.

Table 1. Scoring system used by the 18 raters (6 veterinarians, 6 technicians, and 6 producers) to assess the severity of clinical signs of bovine respiratory disease (BRD) of 524 filmed veal calves

<table>
<thead>
<tr>
<th>Clinical sign</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Nasal discharge</td>
<td>Normal/serous discharge</td>
</tr>
<tr>
<td>Ocular discharge</td>
<td>Normal</td>
</tr>
<tr>
<td>Ear droop or head tilt</td>
<td>Normal</td>
</tr>
<tr>
<td>Abnormal respiration</td>
<td>Normal</td>
</tr>
<tr>
<td>Cough</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 2. Experience years (1 to <5; 5 to <10; +10) of work in calves or veal calves’ health of the 6 veterinarians (V1 to V6), the 6 technicians (T1 to T6), and the 6 producers (P1 to P6) assessing videos of clinical signs associated with bovine respiratory disease of 524 veal calves

<table>
<thead>
<tr>
<th>Group</th>
<th>Rater</th>
<th>Experience (yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veterinarians</td>
<td>V1</td>
<td>5 to &lt;10</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>V4</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>V5</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>V6</td>
<td>1 to &lt;5</td>
</tr>
<tr>
<td>Technicians</td>
<td>T1</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>1 to &lt;5</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>T4</td>
<td>5 to &lt;10</td>
</tr>
<tr>
<td></td>
<td>T5</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>T6</td>
<td>1 to &lt;5</td>
</tr>
<tr>
<td>Producers</td>
<td>P1</td>
<td>1 to &lt;5</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>1 to &lt;5</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>+10</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>5 to &lt;10</td>
</tr>
<tr>
<td></td>
<td>P5</td>
<td>5 to &lt;10</td>
</tr>
<tr>
<td></td>
<td>P6</td>
<td>+10</td>
</tr>
</tbody>
</table>
Briefly, each rater chose a category including the videos of an a priori targeted BRD clinical sign among nasal discharge, ocular discharge, ear droop/head tilt, abnormal respiration, or induced cough. Information concerning the farm, season, and treatment history was not available. Each video was then individually scored by each rater. Before scoring the next video, the score of the current video had to be validated in the software to avoid changing answers. At any time, the participants could save and end their scoring session, and the assessment of the remaining videos could be completed later. The time required to score a video was not recorded and there was no time limit to complete the scoring of all the videos. All raters were aware that the clinical assessment should be compatible with practical use in veal farm units.

### Statistical Analyses

All statistical analyses were performed using SAS v. 9.4 (SAS Institute).

**Power Sample Size Justification of Number of Videos and Raters.** In the absence of data on inter-rater agreement for BRD clinical signs in cattle, we relied on sample size calculation for a traditional Cohen’s kappa (κ) to detect a moderate agreement between raters for each BRD clinical sign (i.e., κ ≥ 0.6; Landis and Koch, 1977). Sample size was consequently determined to obtain the minimum number of videos per BRD clinical sign and the minimum number of raters per group (veterinarians, technicians, or producers) that provide 80% power for testing H₀ = κ < 0.6 versus H₁ = κ ≥ 0.6 at the 5% level of significance. A sample size of 30 videos per BRD clinical sign and 6 raters per group was determined by using the Donner and Eliasziw (1987) method.

**Assessment of Different Scoring Combinations.** Different scoring combinations from the scoring system in Table 1 were assigned to each clinical sign. Briefly, inter-rater agreement was calculated for the original 4-level scale (0/1/2/3) or 3-level scale (0/1/2) and after merging different combinations of adjacent levels to create the 2-level scales (0/1,2,3; 0,1/2,3; 0/1,2; 0/1,2). The following combinations 0,1,2/3 and 0,1/2/3 were not assessed because a limited number of calves were scored 3 by the raters (Table 3). Agreement between each pair of raters was assessed in a k-by-k table, with k being the number of the level scale: 2 or 3.

**Assessment of the Inter-Rater Agreement.** Inter-rater agreement was determined from the k-by-k table using the AgreeStat macro and the function KAPPA from PROC FREQ in SAS. A variety of methods have been described for measuring inter-rater agreement in the literature (Cohen, 1960; Gwet, 2014).

<table>
<thead>
<tr>
<th>Clinical sign</th>
<th>Veterinarians</th>
<th>Technicians</th>
<th>Producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal discharge</td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
<tr>
<td>Ocular discharge</td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
<tr>
<td>Ear droop or head tilt</td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
<tr>
<td>Abnormal respiration</td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
<tr>
<td>Induced cough</td>
<td>Score</td>
<td>V1 V2 V3 V4 V5 V6</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Number of videos of the 524 filmed veal calves for each score assigned to each bovine respiratory disease clinical sign by 18 raters: 6 veterinarians (V1 to V6), 6 technicians (T1 to T6), and 6 producers (P1 to P6)
To date, no single indicator has been considered as the most appropriate measure. We decided, therefore, to estimate inter-rater agreement with the 3 following complementary statistical approaches:

1. Raw percentage of agreement (Pa) was noted as a crude marker of concordant pairs of ratings that does not account for agreement arising from chance alone. It is the most fundamental and intuitive approach. Briefly, Pa is defined as the total number of examinations where agreement is met by 2 raters divided by the total number of videos. Clopper-Pearson 95% confidence intervals (95% CI) were calculated for Pa (Brown et al., 2001). Because of the intervention of multiple raters in this study, the median Pa for each BRD clinical sign was calculated as the median of all Pa values calculated between 2 raters.

2. Agreement beyond chance was assessed using Cohen’s kappa test (κ) between pairs of raters. The kappa reports Pa corrected for agreement due to chance (Pc) corresponding to the probability that a rater classified one video in one category: κ = (Pa-Pc)/(1-Pc) (Cohen, 1960). For the combination including 3 or 4 levels (0/1/2 and 0/1/2/3), the penalty for disagreement was weighted according to the number of categories by which the raters disagreed and the weighted kappa (wκ) was calculated using ordinal weights (Gwet, 2014). The 95% CI were calculated for both κ and wκ using the function KAPPA PROC FREQ in SAS (Fleiss et al., 1969). Because of the intervention of multiple raters in this study, the median κ was calculated as the median of all κ values calculated between 2 raters.

3. Finally, general inter-rater agreement (multiple raters) was evaluated using the Gwet’s agreement coefficient type 1 (AC1). Gwet’s AC1 provides a chance-corrected agreement coefficient, in line with the percentage level of agreement (Gwet, 2014). This parameter allowed estimation of inter-rater agreement of raters in general, rather than for any specific rater pair. It differs from κ mainly in the way the percent chance of agreement is calculated and appears more stable than κ in a homogeneous population (paradox resistant) and consequently is recommended by the authors for use in the medical literature (Wongpakaran et al., 2013). The 95% CI were calculated for AC1 as described by Gwet (2008) using the AgreeStat macro in SAS and by assuming AC1 unconditional to raters or videos.

### Interpretation of the Inter-Rater Agreement.

The most reliable combination for each BRD clinical sign was defined as the combination with the highest inter-rater agreement for at least 2 statistical coefficients (Pa, κ, or AC1). We used the probabilistic statistical approach proposed by Gwet (2014) to interpret the magnitude of the inter-rater agreement for BRD clinical signs. Briefly, for any given agreement coefficient, we first quantified the probability that the extent of agreement falls into each of the benchmark intervals defined by Landis and Koch (1977) (poor agreement for values <0.20; slight agreement for values between 0.21 and 0.40; moderate agreement for values between 0.41 and 0.60; substantial agreement for values between 0.61 and 0.80; and almost perfect agreement for values between 0.81 and 1) (Step 1). For each interval, an interval membership probability (IMP) was calculated. Briefly, it is the probability that a standard normal variate Z falls within the interval, that is:

\[
IMP = P \frac{(X - b)/SE}{Z} \leq (X - a)/SE,\]

where X corresponds to median Pa, median κ or AC1 of a BRD clinical sign, respectively, and SE corresponds to their standard errors; a and b correspond to the lower and upper limits of the intervals, respectively.

We then computed the cumulative probability for each interval starting from the top and going down to the bottom (Step 2). That is, the cumulative probability associated with the top interval is identical to the probability calculated in Step 1 for that interval. For the second interval in descending order, the cumulative probability is the sum of probabilities calculated in Step 1 for the 2 top intervals. The final benchmark level was determined by the interval associated (poor, moderate, substantial, or almost perfect) with the smallest cumulative probability that exceeds 0.95. We therefore concluded with 95% certainty or more that the extent of agreement among raters was poor, moderate, substantial, or almost perfect. The commonly accepted threshold for reliability clinically acceptable was indicated at median κ and AC1 ≥ 0.6, and median PA ≥ 0.75 (Schlageter-Tello et al., 2014).

### Differences Between Veterinarians, Technicians, and Producers.

The inter-rater agreement of the different groups (veterinarians, technicians, and producers) were calculated for each BRD clinical sign. The inter-rater agreement differences between groups were globally estimated by a nonparametric test for dependent samples: the Kruskall–Wallis test. Specific differences between groups (veterinarians vs. technicians, technicians vs. producers, and veterinarians vs.
producers) were estimated by a nonparametric test for dependent samples: the Wilcoxon signed-rank test. All comparisons were performed by using PROC NPAR1WAY from SAS. A $P$-value $< 0.05$ was considered significant to exclude the null hypothesis of equality between groups.

**RESULTS**

**Participants, Videos, and Raters**

Among the 524 calves, 95.8% ($n = 502$) were male. The breeds represented in the population were: Holstein (95%; $n = 497$), Red Holstein (3%; $n = 16$), Jersey (1%; $n = 5$), Ayrshire (1%; $n = 5$), and Brown Swiss (<1%; $n = 1$).

Overall, 149 videos were assessed by 18 raters after the selection process. One video included in the ear droop or head tilt category was not downloaded correctly on the web platform and was not available for the raters. The total video scoring time varied between 2 d and 2 wk for the different raters.

The approximative previous experience in calf or veal calf health of the 6 veterinarians, the 6 technicians, and the 6 producers is presented in Table 2. Seven raters were women (7/18). Most raters (13/18) had more than 5 yr of experience working with veal calves or calf health.

**Overview Inter-Rater Agreement for the Different Scoring Combinations**

Table 3 shows the distribution of severity scores assigned to each BRD clinical sign using the original 3- and 4-level scales by the 18 raters. Despite initially fixing the prevalence of videos to 0.5 (absence-presence of the BRD clinical sign; step 1, Figure 1), the distribution of severity score (1/2 or 1/2/3) obtained was more variable.

Table 4 shows the overall inter-rater agreement of each scoring combination of each BRD clinical sign. Whatever the combination, induced cough was the BRD clinical sign with the highest inter-rater agreement. Abnormal respiration was the BRD clinical sign with the lowest inter-rater agreement.

Within each BRD clinical sign, the 2-level scoring combination 0,1/2,3 (absence-mild/moderate-severe) for nasal discharge, ocular discharge, and ear droop/ head tilt, and 0/1,2 (absence/presence) for abnormal respiration and induced cough had the highest inter-rater agreement.

Table 5 shows the magnitude of the inter-rater BRD clinical sign agreements for the best combination (2 levels). The cumulative probability for each interval was different according to statistical coefficient (median Pa, median $\kappa$, and AC1; Table 5; B). The BRD clinical signs showing a more than 95% probability of belong-
In this study, we estimated inter-rater agreement of BRD clinical signs in veal calves in individual housing (nasal discharge, ocular discharge, ear droop/head tilt; or 0/1,2 (absence/presence) for abnormal respiration and induced cough). Inter-rater agreement was highest (Figure 2).

**Table 5.** Probabilistic statistical approach to interpret the magnitude of the inter-rater bovine respiratory disease (BRD) clinical signs agreements from 254 filmed veal calves, for the 2-level scoring combinations: 0/1,2/3 (absence-mild/moderate-severe) for nasal discharge, ocular discharge, and ear droop/head tilt; or 0/1,2 (absence/presence) for abnormal respiration and induced cough.

<table>
<thead>
<tr>
<th>Probability Benchmark range</th>
<th>Nasal discharge</th>
<th>Ocular discharge</th>
<th>Ear droop/Head tilt</th>
<th>Abnormal respiration</th>
<th>Induced cough</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantified^1 0.8 to 1.0</td>
<td>Almost perfect</td>
<td>0.66 0.02 0.01</td>
<td>0.50 0.03 0.31</td>
<td>0.66 0.15 0.01</td>
<td>0.11 0.01 0</td>
</tr>
<tr>
<td>0.6 to 0.8 Substantial</td>
<td>0.33 0.17 0.84</td>
<td>0.50 0.01 0.69</td>
<td>0.33 0.55 0.49</td>
<td>0.79 0.11 0.43</td>
<td>0.02 0.47 0.16</td>
</tr>
<tr>
<td>0.4 to 0.6 Moderate</td>
<td>0.04 0.43 0.06</td>
<td>0.002 0.38 0</td>
<td>0.02 0.23 0.49</td>
<td>0.11 0.44 0.57</td>
<td>0.00 0.03 0</td>
</tr>
<tr>
<td>0.2 to 0.4 Slight</td>
<td>0.00 0.31 0</td>
<td>0.00 0.31 0</td>
<td>0.02 0.00 0</td>
<td>0.30 0.80 0</td>
<td>0.00 0.00 0</td>
</tr>
<tr>
<td>Poor</td>
<td>0.00 0.07 0</td>
<td>0.00 0.09 0</td>
<td>0.00 0.00 0</td>
<td>0.00 0.00 0</td>
<td>0.00 0.00 0</td>
</tr>
<tr>
<td>Cumulative^2 0.8 to 1.0</td>
<td>Almost perfect</td>
<td>0.66 0.02 0.10</td>
<td>0.50 0.03 0.31</td>
<td>0.66 0.15 0.01</td>
<td>0.11 0.01 0</td>
</tr>
<tr>
<td>0.6 to 0.8 Substantial</td>
<td>0.99 0.19 0.94</td>
<td>1.0 0.20 1.0</td>
<td>0.99 0.70 0.50</td>
<td>0.89 0.12 0.43</td>
<td>0.98 0.93 1.0</td>
</tr>
<tr>
<td>0.4 to 0.6 Moderate</td>
<td>0.99 0.62 1.0</td>
<td>1.0 0.58 1.0</td>
<td>0.99 0.98 0.99</td>
<td>1.0 0.55 1.0</td>
<td>0.98 0.97 1.0</td>
</tr>
<tr>
<td>0.2 to 0.4 Slight</td>
<td>0.99 0.93 1.0</td>
<td>1.0 0.90 1.0</td>
<td>0.99 0.99 1.0</td>
<td>1.0 0.90 1.0</td>
<td>0.98 0.97 1.0</td>
</tr>
<tr>
<td>&lt;0.2 Poor</td>
<td>0.99 0.96 1.0</td>
<td>1.0 0.99 1.0</td>
<td>0.99 0.99 1.0</td>
<td>1.0 1.0 1.0</td>
<td>0.98 0.97 1.0</td>
</tr>
</tbody>
</table>

^1For any given agreement coefficient [median percentage agreement (Pa), median weighted (scoring categories >2) or unweighted (scoring categories <2) kappa (κ), and Gwet’s agreement coefficient type 1 (AC1)], the probability that the extent of agreement falls into each of the benchmark intervals defined by Landis and Koch (1977) was quantified.

^2The cumulative probability for each interval was then computed. The final benchmark level was determined by the interval associated with the smallest cumulative probability that exceeds 0.95. For each interval, interval membership probability (IMP = [(COEFF − b)/SE] ≤ Z ≤ [(COEFF − a)/SE]) was calculated, where COEFF corresponds to Pa, median κ, or AC1 of a BRD clinical sign, respectively; a and b correspond to the lower and upper limit of the intervals, respectively; and Z is a standard normal variate.
just their view angle or limit calf movement to assess the presence and severity of the targeted BRD clinical sign. Additionally, each BRD clinical sign was isolated, which prevented influence from other BRD signs when evaluating a given aspect. Moreover, the raters were informed beforehand what was to be evaluated in the video influencing the mind preset before taking the test. Ideally, the 18 raters should have examined each calf, on site, in a short time period. That seemed technically impossible in veal calf production (limited space, biosecurity, availability of raters). Simultaneous on-farm scoring could also affect individual scoring due to possible interactions between raters. The recording of videos was, therefore, considered the best alternative to standardize the object to be scored (i.e., the same images and video for each scorer).

On each video, the presence and the severity of each BRD clinical sign included in this study was defined based on the 2 clinical scoring systems developed in preweaning dairy calves: the systematic scoring system of WI (McGuirk, 2008) and CA (Love et al., 2014). In contrast with these 2 scoring systems, spontaneous cough could not be systematically assessed because the videos were short. Although often used by operators for BRD detection, this sign is difficult to standardize because spontaneous cough observation cannot be dissociated from the period of observation. For example, a calf coughing every 15 min could be observed for a 1 min period with no coughing, whereas a calf coughing every 2 h would be unlikely to be observed coughing. Therefore, standardization of assessment of the upper airway by pinching the trachea was considered a more reliable and systematic alternative. However, we could not study the variability of stimulating the trachea across operators that could influence our results in real life. Additionally, rectal temperature was measured in our study but its variability between raters was not studied despite it has already been reported (Naylor et al., 2012).

The combination scale of severity differs between both WI and CA. Subjectively, the WI score subdivides each of its clinical signs into 4 levels, whereas CA subdivides them into 2 levels. In our study, we studied the most reliable scale for each BRD clinical sign. We found that a 2-scale combination of clinical signs (0/1/2/3 (absence/mild/moderate/severe) for nasal discharge, ocular discharge, and ear droop or head tilt, or 0/1 (absence/presence) for abnormal respiration and induced cough) was the most reliable scale to assess each BRD clinical sign. It was anticipated that the inclusion of fewer levels of clinical predictors will simplify the assessment of severity and will improve the reliability of each BRD clinical sign, minimizing the possibility of disagreement. Indeed, the multiple level
scale could lead to ambiguous overlap by inexperienced individuals, making it difficult to appropriately classify BRD clinical signs in calves.

Additionally, the simplification from multiple levels to 2 levels has already been reported in dairy cows for improving the lameness scoring system (Schlageter-Tello et al., 2014) as well as in dairy calves’ BRD clinical signs with the CA scoring system (Love et al., 2014). Although the 2-scale combination was the same for induced cough (0 and 1 or 2; presence or absence), abnormal respiration (0/1,2; presence or absence), and ear droop or head tilt (0,1,2,3; absence, mild or moderate, severe) in the CA, the 2-scale combination for nasal discharge and ocular discharge was different with a division of absence/presence (0/1,2) versus absence, mild or moderate, and severe (0,1 or 2, and 3) in our study. However, the statistical methods used in the previous study was different, which could explain the discrepancy with our results.

In accordance with our hypothesis, the inter-rater agreement was different depending on the BRD clinical sign. This finding was in agreement with Buczinski et al. (2016). We found that, whatever the combination, induced cough was the sign with the highest inter-rater agreement. By definition, coughing is a reflex characterized by a forceful and noisy expulsion of air through the glottis (Woolums, 2015). This BRD clinical sign is, therefore, more likely to be detected by the specific movement of the neck and the characteristic sound, visible and audible in videos. Additionally, its severity assessment followed a strict definition without implying interpretation (absence vs. induced single cough vs. induced repeated coughs), making its scoring more objective.

Abnormal respiration was the BRD clinical sign with the lowest inter-rater agreement. In contrast with the induced cough, abnormal respiration indicates an inappropriate degree of breathing efforts based on an assessment of the respiratory rate, rhythm, and character (Woolums, 2015). In our study, we decided to assess abnormal respiration as it was reported in the CA scoring system (i.e., without a specific definition including frequency rate or abnormal dynamic description; Love et al., 2014). Because its assessment is based on different criteria, detecting abnormal respiration needs a minimum of education and definition, which could

Figure 2. Boxplot showing the distribution of kappa value calculated between pairs of raters between the 6 veterinarians (blue), the 6 technicians (red), and the 6 producers (green) for the different clinical signs of bovine respiratory disease. The distribution of the kappa is displayed using 7 values: minimum and maximum (whisker); first quartile (bottom of box), median (horizontal bar), and third quartile (upper of box); mean (+); and outliers (dots). Columns that do not share at least one letter are significantly different (Wilcoxon; P-value <0.05).
explain why abnormal respiration was the sign with the lowest inter-rater agreement. However, it is also possible that abnormal respiration is more difficult to assess by video because of its dynamic characteristic and the contrast between luminosity at farms and the color of the calves, thus decreasing the perception of movement, and thus explaining the low inter-rater agreement observed in our study. Because ambient temperature was recently reported as an important confounder in the BRD detection (Maier et al., 2019), the more objective BRD clinical sign of tachypnea without dyspnea would be of interest in a future study.

The assessment of the other BRD clinical signs (nasal discharge, ocular discharge, ear droop or head tilt) was more precisely defined than abnormal respiration, but it needed more interpretation than induced cough (i.e., it was not only a movement or a noise). This interpretation could explain the variable reliability between raters of these BRD clinical signs. However, ear droop or head tilt showed the highest inter-rater agreement after induced cough. The typically high prevalence of *Mycoplasma bovis* responsible for otitis is commonly observed in veal farms (Radaelli et al., 2008). This characteristic could have made raters more aware of this clinical sign. This result could be different in small herds or other farms where *M. bovis* is less prevalent.

The reliability of individual BRD clinical signs (presence and severity) was never compared between different professions working with veal calves such as veterinarians, technicians, and producers. For 2 statistical coefficients (median Pa and median κ), the veterinarians were more consistent in scoring BRD clinical signs than technicians and producers, except for ear droop or head tilt, where it was more consistently detected by producers. This finding could suggest a lack of standardization between professions in teaching the assessment of BRD clinical signs. This lack of standardization in cattle medicine was described previously by Sischo et al. (2019) who mentioned that veterinarians were infrequently involved in calf-care worker training and that most training occurred either worker-to-worker or manager-to-worker. This difference of reliability in scoring BRD clinical signs could also explain the difference reported in the detection of the disease between veterinarians and producers (Virtala et al., 1999; Olson et al., 2019). Consequently, our results promote the necessity of providing a clear and common definition and formation of each BRD clinical sign to all professions (veterinarians, technicians, and producers) to improve BRD detection, identify appropriate interventions and reduce antimicrobial use.

In this study, we showed that inter-rater agreement varied according to BRD clinical signs. Thus, it is logical to think that if individual BRD clinical signs are not assessed in the same way by 2 different raters, this lack of reliability is also present when we use scoring systems. This low repeatability between raters could explain why current scoring systems lack accuracy (Buczinski et al., 2015; Love et al., 2016). For instance, currently, 2 herds or batches with the same prevalence could have a different proportion of treatment because of the lack of scoring system reliability. Thus, if we want to detect and treat veal calves in individual housing adequately and universally, it is important to implement a scoring system that is used similarly by each operator with the goal of standardizing case definition. Additionally, because confidence interval surrounding any prediction for an individual is substantially wider than for a group, a reliable scoring system would be more accurate if it is applied on a population basis (Hayes et al., 2010).

Interestingly, considering all statistical coefficients, only inter-rater agreements of induced cough and ear droop or head tilt were at least moderate (>0.6) and consequently judged clinically acceptable when applied by raters with different backgrounds. Although a handling of calves will be needed, those 2 BRD clinical signs, in addition to the rectal temperature, could therefore be potentially used to design a reliable scoring system for monitoring and deciding a treatment group. Such a scoring system needs, however, to be developed, validated, and compared with scoring system including all the BRD clinical signs to detect veal calves with bronchopneumonia that require a treatment and not only calves with BRD clinical signs.

The current study has its own limitations. First, the raters included persons working with veal calves and experienced in BRD with minimal training in score assessment. It is, therefore, possible that inter-rater agreement would be different with novice raters or raters inexperienced with veal calves. Second, intrarater agreement was not considered in this study, so it is difficult to know if the difference observed was within a group (veterinarians, technicians, or producers) or because of the inaccuracy of one rater. Additionally, because we generally found better agreement among veterinarians, it is unknown if a more intensive training period with technicians and producers would increase agreement between the latter 2 groups. However, the videos collected during this study could be used to train veterinarians, technicians, and producers to detect BRD clinical signs in the future, as already performed in human medicine (Crawley et al., 2017).

**CONCLUSIONS**

We found that repeatability between raters was highest when BRD clinical signs were assessed with a
2-scale combination (0.1/2.3 (absence, mild/moderate, severe) for nasal discharge, ocular discharge, and ear droop/head tilt and 0/1.2 (absence/presence) for normal respiration and induced cough). Interestingly, among these BRD clinical signs, both induced cough and ear droop or head tilt were the most reliable among raters. In the future, the 2 BRD clinical signs and rectal temperature could be used to design a simple and repeatable screening scoring system. Future research could determine if this discrepancy could be improved through standardized training. Improving the consistency of rating other clinical signs could also help for better day-to-day BRD monitoring.

ACKNOWLEDGMENTS

Funding was received from the Producteurs Bovins du Québec (Longueil, Québec, Canada), Programme Innov’Action Agro-alimentaire (Project IA117743: Amélioration des stratégies diagnostiques de détection des veaux lourds souffrant de pneumonies) du Ministère de l’Agriculture, des Pêcheries et de l’Alimentation du Québec (MAPAQ; Saint Hyacinthe, Québec, Canada), Aliments Pro-lacto inc (Villeroy, Québec, Canada), and FQRNT (Fonds Québécois de Recherche en Nature et Technologie; Montréal, Québec, Canada). The authors thank Mélanie Roche and Paul Mericq (students of the Ecole Veterinaire de Toulouse, France), who participated in the collection of videos during the study. The authors thank Luke Palder from Proofreading Services for the English proofreading. The authors have not stated any conflicts of interest. The video loops used during this study can be obtained upon specific request and further approval by the research team and Délimax veaux lourds Ltée.

REFERENCES


Maier, G. U., J. D. Rowe, T. W. Lehenbauer, B. M. Karle, D. R. Williams, J. D. Champagne, and S. S. Aly. 2019. Development of


ORCIDS

J. Berman o https://orcid.org/0000-0001-7947-760X
D. Francoz o https://orcid.org/0000-0002-1634-0507
S. Dufo a https://orcid.org/0000-0001-6418-0424
S. Buczinski o https://orcid.org/0000-0002-8460-4885