ABSTRACT

The reliability of locomotion scoring is often low, making it unclear how a single gait score should be interpreted. In addition, differences in assessment frequency between longitudinal studies makes it hard to compare results. Our aims were to evaluate how lameness definition and assessment frequency affect measures of lameness incidence. Six dairy farms in British Columbia, Canada, were enrolled, and 262 cows that were sound at dry-off had their locomotion score (LS) assessed weekly from dry-off to calving, using a 1 to 5 scale. Cows were categorized as remaining sound or becoming lame using 3 different case definitions (LAME1: ≥LS3 at least once; LAME2: ≥2 consecutive scores of LS3, or ≥LS4 at least once; and LAME3: ≥3 consecutive scores of LS3, or ≥LS4 at least once). We analyzed the correspondence between the 3 definitions with percent agreement and weighted κ (linear and quadratic weighting). Comparing LAME1 to LAME3 resulted in lower percent agreement (53%) and κ values (linear κw = 0.50; quadratic κw = 0.64) than comparing LAME2 and LAME3 (85%; linear κw = 0.83; quadratic κw = 0.89), indicating that cows scored LS3 twice were likely to be scored LS3 a third time. We also compared the 3 case definitions against trim records from trimmings occurring 90 d or less before calving (n = 117), and used logistic regression models to determine sensitivity, specificity, and positive and negative predictive value. Using the LAME1 criterion resulted in high sensitivity (horn lesions = 0.90; infectious lesions = 0.92) and low specificity (horn = 0.21; infectious = 0.24). We observed higher specificity for LAME2 (horn = 0.62; infectious = 0.66) and LAME3 (horn = 0.71; infectious = 0.77), but LAME2 had higher sensitivity than LAME3 (horn = 0.89 vs. 0.64; infectious = 0.69 vs. 0.64). When evaluating the effects of assessment frequency, we obtained 3 data sets by keeping every, every other, and every third locomotion assessment, and using LAME2 as a case definition. More cows were categorized as lame when assessment frequency increased. Of the cows that were classified as lame when assessed weekly, 72% of the mildly lame, and 33% of the severely lame were classified as sound when assessed every third week. Our results suggest that a single LS3 score should not be used as a criterion for lameness in longitudinal studies. To correctly identify new cases of lameness, dairy cows should be assessed at least every 2 wk.

Key words: disease frequency, gait scoring, hoof pathology, lameness duration

INTRODUCTION

Lameness is one of the greatest welfare challenges for dairy cows, and many cases are associated with claw lesions (Archer et al., 2010). Despite decades of research, lameness remains a common disorder (Bicalho and Oikonomou, 2013), perhaps in part due to the difficulty in correctly identifying lame animals. Intra- and inter-observer reliability estimates often indicate low to moderate consistency when assessing locomotion (reviewed by Schlageter-Tello et al., 2014b), and it is particularly difficult to differentiate between imperfect gait and mild lameness (Schlageter-Tello et al., 2014a). This observer inconsistency introduces uncertainty into measures of prevalence and incidence.

Prevalence estimates (i.e., the proportion of animals lame at a specific time point) from cross-sectional studies have frequently been used to evaluate associations between potential risk factors and lameness (Fauli et al., 1996; Barberg et al., 2007; Solano et al., 2015). Incorrectly classified animals (both sound and lame) increase the risk that true correlations are overlooked, and that spurious relationships are identified.

Cross-sectional studies can conclude only bidirectional associations (Dahoo et al., 2009a), so in recent years more effort has been made to evaluate the temporal associations between risk exposure and lameness using longitudinal studies (Green et al., 2014; Randall et al., 2016). However, in light of poor observer reliability, it is unclear how much weight should be given to single gait scores. To address this issue, some longitudinal studies have used a more stringent definition of lameness

(e.g., 2 consecutive occasions scored as mildly lame, or 1 occasion scored as severely lame; Randall et al., 2015, 2018), but to our knowledge the effect of lameness definition on the diagnosis of new lameness cases has not been systematically evaluated. In addition, no recommendations are currently available for how often gait should be assessed to identify new cases of lameness in longitudinal studies. Assessment intervals vary considerably between lameness studies (e.g., monthly in Frankena et al., 2009; weekly in Randall et al., 2018), making it difficult to compare results.

Using data from dry cows on commercial farms, we aimed to evaluate how lameness definition and assessment frequency affect the identification and classification (mild or severe lameness) of new lameness cases. We predicted that more animals would be categorized as lame when using a less stringent lameness definition and when they were assessed more frequently, resulting in higher incidence estimates. Claw lesions have been imperfectly associated with lameness (Flower and Weary, 2006; Chapinal et al., 2009), and lameness is commonly used to identify cows in need of treatment (Whay, 2002). As such, we also compared the evaluated lameness definitions against trim records from trimmings that occurred 90 d or less before calving.

To evaluate how case definition and assessment frequency affect measures of lameness incidence, we used only data from the 271 parous pregnant animals that were sound at first assessment and kept on-site for the whole dry period. Therefore, the study population was not a representative sample of loose-housed dairy cows, and our results should be interpreted in the context of the study objectives. Each animal was enrolled 9 wk before its estimated calving date. Because of early calving, 33 cows (mean ± standard deviation = 12.6% ± 2.7% of animals per farm) were first assessed 8 wk before their actual calving date. Because later enrollment would have provided insufficient locomotion data, we excluded 9 animals that had their first assessment <8 wk prepartum, leaving 262 cows in the data set. A median of 49 animals (range 9 to 73) was followed per farm.

**Locomotion Assessment**

Before data collection started, 2 observers with previous experience in conducting gait assessments were trained to provide consistent locomotion scores (LS) using an ordinal scale of 1 to 5 (LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move; Flower and Weary, 2006). Initial training in identification of abnormal gait characteristics (head-bob, back arch, shortened track-up, asymmetrical weight bearing, and limp), and overall locomotion scoring was conducted using video recordings of cows walking in a straight line on dry concrete. The initial training was followed by training in live observation, as used in the current study at the University of British Columbia Dairy Education and Research Centre. We evaluated inter-observer reliability after each step of the training protocol (Appendix Tables A1 and A3). Evaluating many animals might exhaust observers and negatively affect assessment accuracy (Schlageter-Tello et al., 2015), so we controlled for the number of animals used for the live reliability assessment.

Inter-observer reliability for ordinal gait assessment was evaluated using weighted κ (Cohen, 1968) with linear and quadratic weighting (Vanbelle, 2016). Using the framework of Landis and Koch (1977), the κ values indicated substantial to almost perfect agreement [linear weighting κw = 0.76, 95% confidence interval (CI) 0.53–0.99; quadratic weighting κw = 0.84, 95% CI 0.69–1.00] for video assessments, and moderate agreement (linear κw = 0.52, 95% CI 0.24–0.80; quadratic κw = 0.57, 95% CI 0.28–0.87) when evaluating live scoring reliability.

To evaluate the effects of prevalence and observer bias on the reliability measures, we dichotomized locomotion data as sound (≤LS3) or clinically lame (≥LS3; Appendix Tables A2 and A4), and calculated the prevalence and bias indexes for video and live training.
as described by Byrt et al. (1993). Both indexes can take values between −1 and 1; a prevalence index of 0 indicates that equal numbers of animals were classified as sound and lame, and a bias index of 0 indicates that no systematic bias existed between observers. For video assessment, the prevalence index was 0.1 and the bias index was 0; for live evaluation, the prevalence index was 0.3 and the bias index was −0.1.

We reassessed reliability for live scoring after data collection was completed (linear κw = 0.49, 95% CI 0.19–0.80; quadratic κw = 0.55, 95% CI 0.23–0.87; Appendix Table A5), and found prevalence and bias index values of 0.2 and 0, respectively (Appendix Table A6). The similar results before and after the study period indicate that the observers maintained reliability and did not develop systematic bias over the study period.

During the study, locomotion was scored weekly during pen walks, when unrestricted animals walked on concrete while being observed directly and obliquely from behind by one of the observers. Farmers were told about animals with a score ≥LS3, but treatment was initiated at the discretion of the farm manager.

**Claw Lesion Recording**

Farms were visited every 2 to 10 wk by the same trimming company used to recruit the farms. The cows were trimmed by 1 of 3 professional trimmers, who also recorded claw lesions in accordance with the Alberta Dairy Hoof Health Project’s Lesion Severity Scoring Guide (www.hoofhealth.ca/Section5/LesionSeverityGuide_v0.5.pdf), using Hoof Supervisor System software (KS Dairy Consulting, Inc., Dresser, WI). On 5 of the enrolled farms, the animals were routinely trimmed before dry-off, and on the sixth farm animals were selected for trimming based on toe length. After the on-farm data collection was completed, we downloaded records for trimmings that occurred in 2017 and extracted lesion records from trimmings that occurred 90 d or less before calving, including during the dry period. Of 262 enrolled cows, only 117 had trimming data within this period. Each animal was classified as having or not having noninfectious [sole hemorrhage (n = 1), sole ulcer (n = 3), white line lesion (n = 4), toe ulcer (n = 2); 1 animal had >1 type of claw horn lesion] and infectious claw lesions [digital dermatitis (n = 22), interdigital dermatitis (n = 6); 2 animals had both types of infectious lesions].

**Data Handling and Statistical Analyses**

**Lameness Definition.** To evaluate the effects of case definition on the identification of new lameness cases, clinical lameness was defined as follows: (1) scored LS3 or higher on at least 1 occasion (LAME1); (2) scored LS3 on at least 2 consecutive occasions or scored LS4 or higher at least once (LAME2); or (3) scored LS3 on at least 3 consecutive occasions or scored LS4 or higher at least once (LAME3). We justified the use of 1 score of ≥LS4 by the relative ease of identifying animals with severely restricted ability to move (Schlageter-Tello et al., 2014a), making it unlikely that sound animals would be scored LS4 or higher. We treated LAME3 as the gold standard for clinical lameness and used ≥LS4 to define severe lameness (Flower and Weary, 2006).

**Assessment Frequency.** When evaluating the effects of assessment frequency, we retained the results of every assessment (ASSM1), every other assessment (ASSM2), or every third assessment (ASSM3) for each animal; the first assessment was the same for all 3 data sets. For cows that were gait scored for the first time 9 wk before calving (87% of enrolled animals), 4 observations remained in ASSM3; animals first observed 8 wk prepartum (13%) were scored 3 times. After creating the 3 data sets, we used the case definition found to have the best test characteristics to categorize cows as sound or lame, treating ASSM1 as the gold standard.

**Statistical Analyses.** All data analyses were performed in R 3.4.4 (RStudio Team, 2016; Wickham et al., 2017; R Core Team, 2018; Wickham and Henry, 2018). Graphing was performed using the ggplot2 package (Wickham, 2009). Mixed logistic regression models with farm as random intercept were fitted with maximum likelihood using the lm4 package (Bates et al., 2015) to estimate sensitivity (proportion of true positives that were correctly identified by the test), specificity (proportion of true negatives correctly identified by the test), positive predictive value (PPV; probability of correct positive test), and negative predictive value (NPV; probability of correct negative test) as described by Coughlin et al. (1992) to evaluate the effect of assessment frequency. We evaluated the uncertainty of the estimates with exact binomial 95% confidence intervals (Dahoo et al., 2009b). We calculated percent agreement from contingency tables. Because the 3 lameness criteria were nested, it was not possible to estimate their relative test attributes with logistic regression due to complete data separation; as such, we split the data per farm when evaluating correspondence between the 3 case definitions, and we calculated sensitivity, specificity, PPV, NPV, and percent agreement from contingency tables per farm. However, we used mixed logistic regression models when comparing the different lameness criteria against recorded claw lesions.

We evaluated reliability in the classification of lameness severity (sound, mildly lame, severely lame) depending on lameness definition and assessment fre-
frequency using weighted κ (linear and quadratic weighting; Vanbelle, 2016) and the irr (Gamer et al., 2012) and rel (Lo Martire, 2017) packages. The κ statistic assumes that the data are independent, which has generally been interpreted as observers having no knowledge of previous assessments (Sim and Wright, 2005), and that repeated assessments of the same subject must be accounted for (Haley and Osberg, 1989). To evaluate whether dependencies in our data affected the reliability estimates, we obtained κ values first for all animals, and then separately for the 3 largest herds (n = 50–73).

Because lameness incidence was calculated from a closed population, we used incidence proportion as the outcome (number of cows becoming lame/total number of cows; Dahoo et al., 2009b). We tested differences in farm incidence depending on lameness definition and assessment frequency using Wilcoxon signed rank tests. The low number of enrolled animals in 1 of the farms (n = 9) increased uncertainty in the incidence estimate for this farm; therefore, we ran all analyses with and without animals from this farm. Excluding the data for this farm caused no substantial differences in the results, so analyses using the full data set are presented here. The data sets and accompanying R script are available at https://doi.org/10.5683/SP2/9XRRSJ.

RESULTS AND DISCUSSION

Locomotion Assessment

The high weighted κ values after video training suggested that the observers could differentiate between gait scores. The lower values for the linear-weighted κ statistic compared with the quadratic-weighted κ statistic indicate that disagreements generally consisted of 1 LS unit (Vanbelle, 2016). The consistently low bias index values suggested that the lower weighted κ values for live scoring were not owing to systematic differences in how the observers scored locomotion, but rather that it was difficult to differentiate between LS2 and LS3 (Appendix Table A4 and Table A6; Vach, 2005). This finding aligned with that of Schlager-Tello et al. (2015), who reported the lowest intra-observer percent agreement for LS2 and LS3 scores. Taken together, these results emphasize that the interpretation of longitudinal locomotion data should account for observer inconsistency.

On 10 occasions, locomotion assessment could not be performed (0.4% of all occasions); on 8 occasions, the cows were calving during the visit and on 2 occasions, the cows were not found in their home pens. No animal lacked more than 1 locomotion score. The percentages of LS1, LS2, LS3, LS4, and LS5 were 0, 69.5, 26.5, 4, and 0%, respectively; 18, 60, and 22% of the animals had a maximum score of LS2, LS3, and LS4, respectively. The reason that no animal was scored as having a perfect gait may have been that the home pens did not provide optimal conditions for walking freely (other animals present, manure-contaminated concrete; Telezhenko et al., 2017). It was common for locomotion scores to change from one visit to the next; approximately one-third of the scores changed between consecutive assessments. In agreement with Reader et al. (2011), scores were much more likely to change by a single unit (e.g., from LS2 to LS3, or LS3 to LS4; 94% of all changes) than by a score of 2 or more (e.g., from LS2 to LS4). Some of these changes could be explained by inconsistency within and between observers (Figure 1A and C), as well as by animals transitioning between soundness and lameness, and vice versa (Figure 1D and E).

Effect of Lameness Definition

The number of cows categorized as becoming lame differed depending on the definition used [LAME1 n = 215 (82% of enrolled animals); LAME2 n = 131 (50%); LAME3 n = 93 (35%)]. We attributed these differences to animals being considered sound when using the LAME3 criteria (n = 169), while being classified as lame when LAME1 (122 of 169 animals) or LAME2 (38 of 169 animals) criteria were used. The high proportion of animals categorized as lame when using LAME1 suggests that the lameness incidence was overestimated, especially considering that 44 of the 158 lame (28%) LAME1 animals received a score of LS3 on only 1 occasion.

The higher specificity, PPV, and percent agreement (Table 1) found when comparing LAME2 and LAME3 suggests that animals scored ≥LS3 on 2 consecutive occasions were likely to be scored ≥LS3 a third time. This finding was also reflected by weighted κ values indicating almost perfect agreement (linear weighting κw = 0.83, 95% CI 0.77–0.88; quadratic weighting κw = 0.89; 95% CI 0.85–0.93) between LAME2 and LAME3, versus only moderate to substantial agreement (linear κw = 0.50, 95% CI 0.43–0.57; quadratic κw = 0.64, 95% CI 0.58–0.70) between LAME1 and LAME3. We found the same patterns when evaluating reliability measures separately for the 3 largest farms, with better agreement between LAME2 and LAME3 (linear κw = 0.77–0.87; quadratic κw = 0.84–0.92), than when comparing LAME1 to LAME3 (linear κw = 0.36–0.61; quadratic: κw = 0.47–0.73) for all 3 farms. Farm incidence measures were higher when using LAME1 (average farm incidence = 81%; Z = 21; P = 0.03) and LAME2 (average farm incidence = 48%; Z = 15; P = 0.06), compared with LAME3 (average farm incidence = 35%).
Our findings demonstrate that assigning equal weight to occasional and sequential LS3 scores introduces noise when identifying new cases of lameness (Figure 1A). It is possible that if we had performed locomotion scoring in more controlled conditions (i.e., not on commercial farms), the number of false positives would have been lower. Although we did not systematically assess movement when transitioning from lying to standing, we observed that heavy animals in late gestation were often visibly stiff directly after rising; this may have resulted in a higher rate of misclassification of lameness for the cows used in this study.

The low agreement between LAME1 and LAME3 suggests that lameness prevalence estimates contain uncertainty and should be viewed with caution. It is possible that the reported divergence in estimates of lameness occurred because the nested lameness criteria led to complete data separation when analyzed using mixed logistic regression, values for lameness criteria were obtained per farm (n = 6) from contingency tables. We used mixed logistic regression models (farm included as random intercept) to estimate values for assessment frequency. We used LAME3 as the gold standard for lameness, and compared it against LAME1 and LAME2. We used ASSM1 as the gold standard for assessment frequency, and compared it against ASSM2 and ASSM3; for all assessment intervals, we used LAME2 as the lameness criteria. Percent agreement was calculated from contingency tables. Lameness scores were as follows: LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move.

Observed scores were as follows: LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move.

| Item | Sensitivity (95% CI)² | Specificity (95% CI)² | PPV (95% CI)² | NPV (95% CI)² | Agreement, %
|------|----------------------|----------------------|---------------|---------------|----------------
| Lameness criteria (n = 262)³ | | | | | |
| LAME1 | 0.13−0.40⁴ | 0.24−0.57⁴ | — | — | 43−67 |
| LAME2 | — | 0.68−1.00⁴ | 0.50−1.00⁴ | — | |
| Assessment frequency (n = 262) | | | | | |
| ASSM2 | 0.68 (0.62−0.74) | 0.95 (0.92−0.97) | 0.93 (0.89−0.96) | 0.75 (0.69−0.80) | 79−100 |
| ASSM3 | 0.44 (0.39−0.51) | 0.98 (0.95−0.99) | 0.95 (0.92−0.97) | 0.64 (0.58−0.70) | 71 |

1LAME1 = ≥LS3 on ≥1 occasion; LAME2 = LS3 on ≥2 consecutive occasions or ≥LS4 on ≥1 occasion; LAME3 = LS3 on ≥3 consecutive occasions or ≥LS4 on ≥1 occasion; ASSM1 = weekly gait assessments; ASSM2 = gait assessments every 2 wk; ASSM3 = gait assessments every 3 wk. Because the nested lameness criteria led to complete data separation when analyzed using mixed logistic regression, values for lameness criteria were obtained per farm (n = 6) from contingency tables. We used mixed logistic regression models (farm included as random intercept) to estimate values for assessment frequency. We used LAME3 as the gold standard for lameness, and compared it against LAME1 and LAME2. We used ASSM1 as the gold standard for assessment frequency, and compared it against ASSM2 and ASSM3; for all assessment intervals, we used LAME2 as the lameness criteria. Percent agreement was calculated from contingency tables. Lameness scores were as follows: LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move. LS = locomotion score.

2Exact binomial confidence intervals.

3Because of nested lameness criteria, sensitivity and NPV values had to equal 1.00. Because these values were an effect of our definitions rather than true test characteristics, the values are not shown.

4The range represents the lowest and highest values for the farms.

Figure 1. Weekly locomotion scores (LS) for 6 different animals. The points illustrate recorded gait scores, and the lines represent interpolations of gait change between assessments. When one ≥LS3 was used to define lameness, cow A was considered a new lameness case, contributing to farm incidence. With decreasing assessment frequency, it became more difficult to identify all lame animals; cow B was categorized as sound when evaluated every 2 wk despite multiple scores of LS4, and cow C when assessed every 3 wk (mild lameness was defined as ≥2 consecutive LS3 when evaluating assessment frequency). The presence of isolated LS2 scores at the beginning (D) and end (E) of severe lameness events suggests that it is difficult to distinguish between LS2 and LS3 in marginally lame animals, which raises the question of how to interpret the trajectory for cow F. Lameness scores were as follows: LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move.
lameness prevalence between farmers and trained assessors (e.g., Espejo et al., 2006; Leach et al., 2010) is not because assessors are more skilled in identifying mildly lame animals (e.g., Alawneh et al., 2012), but because farmers base their evaluations on multiple observations. In a qualitative study, Horseman et al. (2014) reported that many of the enrolled farmers (n = 12) were able to detect mild lameness, but referred to this as impaired mobility. Leach et al. (2012) reported that some lame cases seem to self-cure, which may explain why farmers consider only LS4 and LS5 as “true” lameness. Although we used LAME3 as the gold standard in our analyses, the duration of some true lameness cases might have been too short to be correctly identified under the conditions of this study (weekly gait assessments). Given the high agreement and reliability between LAME2 and LAME3, we suggest that 2 consecutive LS3 scores is a reasonable cutoff for new cases of mild lameness in longitudinal research.

Relation Between Lameness Definition and Claw Lesions

Table 2 shows the percentage agreement, sensitivity, specificity, PPV, and NPV when using LAME1, LAME2, and LAME3 to identify animals with noninfectious and infectious claw lesions. The high sensitivity we found for LAME1 suggests that most animals with claw lesions had at least a brief period of altered gait during the dry period. However, the low specificity and PPV indicate that short-term gait impairment was also observed in many of the animals that did not have claw lesions at trimming, suggesting that the LAME1 criterion may introduce bias in the evaluation of longitudinal locomotion data.

When we used LAME2 and LAME3 definitions, the number of false positives decreased, leading to higher specificity and percent agreement for both noninfectious and infectious lesions (Table 2). When comparing LAME2 and LAME3, LAME2 had higher sensitivity (noninfectious 0.89 vs. 0.64; infectious 0.69 vs. 0.64), and lower specificity (noninfectious 0.62 vs. 0.71; infectious 0.66 vs. 0.77). However, we also obtained low PPV when LAME2 and LAME3 were used as lameness criteria, possibly because painful lesions at the level of the corium were not visible on the sole surface at the time of trimming. All case definitions had high NPV, indicating that cows that never had a gait score ≥LS3 were unlikely to have lesions at trimming.

Effects of Assessment Frequency

We used the LAME2 criteria when evaluating the effect of assessment frequency. Fewer animals were categorized as becoming lame when assessment frequency decreased [ASSM1 n = 131 (50% of enrolled animals); ASSM2 n = 96 (37%); ASSM3 n = 62 (24%)]. The increasing difficulty in correctly identifying new lameness cases when assessment occurred less frequently was reflected in the higher sensitivity and NPV for ASSM2 compared with ASSM3 when evaluated against weekly assessments (Table 1).

The weighted κ values showed substantial agreement between ASSM1 and ASSM2 (linear κw = 0.68, 95% CI 0.60–0.76; quadratic κw = 0.75, 95% CI 0.67–0.82), and moderate agreement between ASSM1 and ASSM3 (linear κw = 0.48, 95% CI 0.39–0.57; quadratic κw = 0.55, 95% CI 0.45–0.65; Landis and Koch, 1977). Similar results were obtained when evaluating reliability measures for the 3 largest farms, with higher agreement.
between weekly and fortnightly assessments (linear $\kappa_w = 0.63–0.78$; quadratic $\kappa_w = 0.73–0.85$), than when comparing weekly assessments with assessments every 3 wk (linear $\kappa_w = 0.12–0.60$; quadratic $\kappa_w = 0.12–0.68$).

A large percentage of animals categorized as lame when assessed weekly was classified as sound when assessed every other week (47% of mildly lame and 12% of severely lame animals), and every third week (72% of mildly lame and 33% of severely lame animals). Approximately 20% of the animals categorized as severely lame when assessed weekly were classified as mildly lame when assessed less frequently. Farm incidence estimates were higher when animals were scored every week (average farm incidence = 48%), compared with every other week (average farm incidence = 36%; $Z = 15; P = 0.06$), and every third week (average farm incidence = 25%; $Z = 15; P = 0.06$).

### Lameness Duration

To explore why lameness classification differed between ASSM1, ASSM2, and ASSM3, we determined the number of weeks an animal was categorized as lame from ASSM1, using 2 consecutive scores of LS2 as the definition of being cured. Because the data collection ended at calving, the lameness duration for animals classified as lame on the last assessment was unknown. We found that 12% of the animals categorized as lame for $\leq 2$ visits were lame when censored; the corresponding value for animals lame for $\geq 4$ visits was 78%. Of the animals that changed category from lame to sound when assessed less frequently, 69% (ASSM2) and 42% (ASSM3) were lame for $\leq 2$ wk, and 21% (ASSM2) and 49% (ASSM3) were lame for $\geq 4$ wk.

Figure 2 shows a histogram of the number of weeks classified as lame for mildly and severely lame animals. The markedly bimodal distribution of lameness duration for mildly lame animals suggests that several animals were incorrectly classified as lame when assessed weekly, or that this distribution represents 2 populations of animals with different underlying causes of lameness. In our study, 35% of the animals with infectious claw lesions (9 of 26 animals) were scored $\geq$LS3 for $\leq 1$ wk, but the same was true for just 11% of the animals with noninfectious claw lesions (1 of 9 animals). Because lame cows were reported to farmers after every visit, it is possible that treatment was initiated shortly after the onset of lameness, at least on some farms. Early treatment may have shortened the duration of readily treatable lameness cases, resulting in the bimodal distribution of lameness duration.

### Lameness Duration Diagram

**Figure 2.** Number of weeks that cows were lame during the dry period, for mildly and severely lame cows. Mild lameness was defined as a locomotion score (LS) of LS3 for $\geq 2$ consecutive weeks, and severe lameness was defined as at least 1 score of $\geq$LS4. Lameness scores were as follows: LS1 to LS2 = freely able to move; LS3 = compromised locomotion; LS4 to LS5 = severely restricted ability to move.
gait assessments, suggesting that the same challenges of interpreting longitudinal locomotion data are present also after calving.

Implications for Farm Management

A practical limitation of the current study is that we assessed locomotion on a weekly basis. Although this is a quite frequent sampling regimen for a longitudinal lameness study, it differs considerably from the daily (albeit unsystematic) assessments that farmers self-reported in a study conducted by Horseman et al. (2014). Such short assessment intervals may enable farmers to use information from multiple gait assessments to more reliably identify mildly lame animals while still being able to provide treatment within a reasonable time period. Further studies should evaluate what case definition would be most useful when cows can be assessed multiple times per day. Because it is unlikely that animals with severely impaired gait are actually sound, the examination and treatment of cows that easily can be identified as lame should be performed as soon as possible after first detection (Leach et al., 2012).

Some animal groups are not observed walking during daily work tasks (e.g., dry cows); we suggest that these animals should be evaluated for lameness at least every other week. This procedure may also be valuable in dairy systems with less intensive everyday human–animal contact. Given the effort required to conduct gait assessments this frequently, we urge work on the development of automated locomotion scoring systems suitable for use on commercial farms.

CONCLUSIONS

Estimates of lameness incidence increased when a less stringent lameness definition was used, and when assessment frequency increased. Our results suggest that occasional observations of LS3 should not be used as a criterion for mild lameness in longitudinal studies. Decreased assessment frequency made it more difficult to identify lame animals, suggesting that dairy cows should be assessed at least every other week to adequately identify new cases of lameness.

ACKNOWLEDGMENTS

We thank the participating farmers and their staff for their help, and for patiently answering all our questions. We thank Samantha Jung, Paige McDonald, and Wali Sahar from the Animal Welfare Program for their assistance with data collection, and Martin Stigge for software support and help with R script development. During this project, HKE was supported by...
REFERENCES


Eriksson et al.: FACTORS AFFECTING LAMENESS INCIDENCE ESTIMATES


ORCIDs

Ruan R. Daros. https://orcid.org/0000-0003-2331-1648
Marina A. G. von Keyserlingk. https://orcid.org/0000-0002-1427-3152
Daniel M. Weary. https://orcid.org/0000-0002-0917-3982
In this appendix, we present complementary material related to the evaluation of inter-observer reliability measures for locomotion assessment. During training of the 2 observers who assessed locomotion, we used recorded videos and live scoring. We evaluated observer reliability after each step in the training, and after completing data collection. Tables A1 and A2 illustrate interobserver agreement for the ordinal scale and classification of animals as sound or clinically lame from video. Tables A3 and A4 summarize the data from the pre-study evaluation of live locomotion assessment, and Tables A5 and A6 present the data from the reassessment of observer reliability after study completion. We used data in Tables A2, A4, and A6 when calculating prevalence and bias indices.

**Table A1. Inter-observer agreement for gait scoring from video (ordinal scale 1 to 5)**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table A2. Inter-observer agreement for lameness categorization (sound or lame) from video**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Sound</th>
<th>Lame</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>10</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Sound</td>
<td>9</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>Lame</td>
<td>11</td>
<td>9</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table A3. Inter-observer agreement for pre-study live gait scoring (ordinal scale 1 to 5)**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>27</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>40</td>
</tr>
</tbody>
</table>

**Table A4. Inter-observer agreement for pre-study live lameness categorization (sound or lame)**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Sound</th>
<th>Lame</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>20</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Sound</td>
<td>7</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>13</td>
<td>40</td>
</tr>
</tbody>
</table>

**Table A5. Inter-observer agreement for live gait scoring after study completion (ordinal scale 1 to 5)**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>5</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>22</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>37</td>
</tr>
</tbody>
</table>

**Table A6. Inter-observer agreement for live lameness categorization after study completion (sound or lame)**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>Sound</th>
<th>Lame</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 2</td>
<td>17</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Sound</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>15</td>
<td>37</td>
</tr>
</tbody>
</table>