Journal of Dairy Research

cambridge.org/dar

How reliable is visual lameness scoring? Assessing human label variability for use in automated detection systems

Konstantina Linardopoulou[1], Lorenzo Viora[1], George King[1], Julien Le Kernec[2] and Nicholas N Jonsson[1]

[1]School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, Glasgow, UK and [2]James Watt School of Engineering, University of Glasgow, Glasgow, UK

Research Article

Cite this article: Linardopoulou K, Viora L, King G, Le Kernec J and Jonsson NN (2026) How reliable is visual lameness scoring? Assessing human label variability for use in automated detection systems. *Journal of Dairy Research* **93**, 258–264. https://doi.org/10.1017/S0022029926102180

Received: 21 July 2025
Revised: 18 November 2025
Accepted: 11 January 2026
First published online: 9 March 2026

Keywords:
animal behaviour; animal science; animal welfare; dairy management; lameness

Corresponding author:
Konstantina Linardopoulou,
Email: konstantina.linardopoulou@glasgow.ac.uk

Abstract

Visual mobility scoring to detect lame dairy cattle can be subjective and inconsistent. This study assessed the reliability of visual mobility scores from multiple assessors, using different scoring methods (live vs. video) and experience levels to evaluate their influence on label quality for machine learning applications. We gathered data from two farms using the AHDB 4-point mobility scale and a simplified post-hoc dichotomised version, with both live and video assessments. Substantial within- and between-assessor variation was seen in scores, particularly for scores 0 and 1 (consistent with normal and slightly abnormal gaits, respectively). Assessors showed only fair (weighted kappa ≈ 0.33) score consistency when they scored the same animal in different ways (live vs. video). Post-hoc simplification of the four-level scores to a dichotomous score improved agreement but reduced granularity. Assessor experience had limited influence on agreement levels ($P > 0.05$), and increased video viewing frequency during the assessment process was associated with lower inter-assessor agreement (probability estimate = −0.49, $P = 0.005$), suggesting higher uncertainty in ambiguous cases. Qualitative feedback from assessor comments revealed that the speed of the animal affected their scoring decisions ($\beta = -1.92$, $P = 0.007$). These results highlight the difficulties in using subjective human scores as labels for machine learning training. To improve automatic lameness detection in dairy cattle, we need strategies to reduce this variation and use more definitive labels.

Early diagnosis of mobility problems in dairy cows is crucial for keeping the animals healthy and productive, thereby supporting a sustainable production system. When a cow becomes lame, it can lead to weight loss and poor body condition (Huxley, 2013), pain, and discomfort (Shearer *et al.*, 2013). If treatment is delayed, it can result in lower milk production, poor fertility, higher culling rates, greater greenhouse gas emissions intensity per unit of milk produced, and increased veterinary costs for farmers (Green *et al.*, 2002; Amory *et al.* 2008; Mostert *et al.*, 2018; Omontese *et al.*, 2020). Early detection of lameness in dairy animals allows for prompt action, preventing the development of more complex issues.

With technological advancement, interest in automating lameness detection is growing. Radar, computer vision, and other sensor technologies-based systems are expected to provide more precise and consistent evaluations compared to traditional approaches (Schlageter-Tello *et al.*, 2014; Busin *et al.*, 2019; Nektarios *et al.*, 2024). Computer-based systems provide the capability for continuous, real-time observation of animals with very little human interaction, cutting costs and subjectivity that come with manual visual observations. Nevertheless, these systems still face challenges related to data reliability, particularly because they often rely on manually labelled data as the ground truth for training the associated machine learning models (Busin *et al.*, 2019).

Currently, lameness detection relies on visual mobility scoring systems, like the AHDB 4-level mobility system (Agriculture and Horticulture Development Board, 2025). In this system, trained assessors visually appraise animals and assign a lameness score. Although such methods are widely used, they are also subjective and inconsistent (Engel *et al.*, 2003; Nielsen *et al.*, 2014; Schlageter-Tello *et al.*, 2014). The variation between different assessors (inter-assessor), and between the same assessor at different times (intra-assessor), can add significant noise to the data (Kahneman *et al.*, 2016), undermining the reliability of visual scoring as a basis for training algorithms.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Hannah Dairy Research Foundation. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

The subjective and unreliable nature of visual scoring has prompted us to investigate how best to use human-labelled data to train machine learning models. Particularly, we wished to assess the homogeneity of such labels and whether variables such as the experience of the assessor, the method used for scoring (live or video), and data quality in general influenced automated system performance. This study aims to evaluate these factors, and their impact on the quality of labels that are employed to train ML models for early signs of lameness detection.

Materials and methods

This study complied with the University of Glasgow's ethical guidelines and received approval from the local ethics committee (licence EA06 19), although it did not involve any procedures regulated under the Animals (Scientific Procedures) Act 1986 (UK Public General Acts, 1986).

Due to COVID-19 restrictions, we utilised live and video methods for mobility scoring and could not maintain the same group of assessors. Consequently, we adapted our experimental design to collect data while managing limited farm access (Table 1).

Farm A – live and video-based scoring

In this study, five and seven assessors scored 49 cows on-site during Visit 1 and 52 cows during Visit 2 using the AHDB mobility scoring system. The scores were later re-analysed using a modified version of the system in post hoc analysis. Farm A is a commercial dairy farm in Scotland, with Holstein–Friesian cows housed in a free-stall barn and walked regularly on concrete floors and rubber-matted alleyways at the milking parlour and its exit.

Cows were walked along their everyday routine alley, and assessors evaluated each animal's gait from a lateral vantage point, standing approximately 5 m away from the walking cows. All assessors were positioned in the same area, with an unobstructed lateral view of the animals as they walked on the same surface and under identical environmental conditions. The AHDB scoring system included four ordinal categories (0: sound, 1: imperfect mobility, 2: impaired mobility, 3: severely impaired mobility), and all assessors received brief refresher instructions before the session, and they scored six to eight videos together.

Assessor experience levels varied (Table 2): from experts, who were accredited registered mobility scorers (RoMS) assessing mobility routinely as part of their professional work, to experienced assessors, who were veterinarians with over 10 years of experience and who regularly included mobility scoring as part of farm visits (not RoMS-accredited). Moderately experienced assessors had experience in mobility assessments through research or farm work but did not perform them routinely. Novice assessors who participated in the study had limited or no previous experience with mobility assessment. Video recordings were taken during Visit 1 only to allow further video-based scoring in follow-up studies.

Video recordings were shared with six assessors (Table 2) for independent scoring using the AHDB mobility scoring system. Each assessor reviewed the videos independently, with no limit imposed on the number of times a video could be watched. Along with the scores, metadata were collected, including self-reported confidence in each score (Yes/No), comments regarding any potential obstructions to scoring (such as unclear video footage), and how many times each video was viewed before making a decision. This study enabled the exploration of scoring consistency in a remote, asynchronous setting.

Farm B – remote video-based scoring with metadata collection

A separate set of Holstein–Friesian cows ($n = 69$) was scored using video recordings obtained from Farm B, a commercial dairy farm in central Scotland. Videos were recorded from the lateral view as cows walked through a designated passageway. No live on-site scoring was conducted. Eight assessors independently scored the videos using the same AHDB mobility scoring system as in Farm A. The same metadata (confidence, obstructions, number of views) as in Farm A were also collected.

Data processing

Lameness scores from all studies were dichotomised post hoc into two categories: not-lame (sound) (0) and lame (1). Cows classified as "sound" had scores of 0 or 1, while cows classified as "lame" had scores of 2 or 3, following the AHDB scoring system. Each individual score was treated as an independent data point.

Inter-assessor agreement for all datasets (AHDB & dichotomised AHDB) was assessed using weighted kappa (κ_w) to evaluate consistency across assessors. Intra-assessor agreement was also calculated for repeated scoring to compare the scores collected on-site with those recorded on video. Assessors' self-reported confidence levels were examined alongside the number of video views they took before finalising their scores. Descriptive statistics were used to summarise confidence and number of views, and correlations between these variables and scoring consistency were explored.

To investigate factors potentially influencing agreement, descriptive statistics were used to summarise assessors' confidence levels and viewing behaviours. We examined whether assessor experience level (Novice, Moderate, Experienced, Expert) was associated with differences in inter-assessor agreement. Agreement scores (weighted kappa values) were compared between experience groups for each dataset (Farm A Visit 1, Visit 2, Farm A Video, and Farm B Video) using Kruskal–Wallis rank-sum tests. Where we found differences, we carried out post hoc pairwise comparisons and adjusted the results using Bonferroni correction to account for multiple tests.

We used a generalized linear mixed-effects model (GLMM) to see if average confidence and viewing frequency could predict how often assessors agreed. The outcome was whether assessors agreed (1 = agreement, 0 = disagreement) while scoring the same video. We included the average certainty score (binary, 0 = uncertain, 1 = certain) and the average number of views per video (continuous) as fixed effects. To account for differences in individual scoring behaviour, we added assessor identity as a random intercept. We applied a binomial distribution with a logit link function for the model.

To further explore how qualitative comments related to scoring behaviour, we fitted a proportional odds logistic regression model with the AHDB mobility score as the ordinal outcome. Thematic categories derived from assessor comments (Speed Issues, Posture and Gait Observations, Visibility/Video Issues, Uncertainty in Diagnosis, External Physical Condition, and Other Technical or Personal Notes) were included as fixed predictors to examine whether specific types of comments were associated with higher or lower mobility scores.

https://doi.org/10.1017/S0022029926102180 Published online by Cambridge University Press

Table 1. Study design for two farms (A and B), including the number of visits, cows assessed, scoring methods (on-site and video), number of assessors, scoring systems and modes used, and types of metadata collected during video-based assessments

Farm	A				B		
Visits	2				1		
Number of cows scored	Visit 1 → 49			Visit 2 → 52			69
On-site scoring	Yes			Yes			No
Video scorings	Yes			No			Yes
Assessors	6 (on-site)	7 (video)		5	8		
Scoring System	AHDB + post hoc modifications				AHDB + post hoc modifications		
Scoring Mode	Individual				Individual		
Metadata Collected	Assessor experience	Assessor confidence (video)	Number of video views	Assessor experience	Assessor experience	Assessor confidence	Number of video views

Table 2. Thirteen assessors with differences in experience and occupation scored the animals in the studies. Only two assessors (R1 and R2) participated in all four evaluations, and another five assessed the videos from both Farm A and Farm B

Assessors	Occupation	Experience in mobility scoring	Assessments			
			Farm A Visit 1	Farm A Visit 2	Farm A Video	Farm B Video
R1	Bovine Veterinarian	Experienced	✓	✓	✓	✓
R2	Bovine Veterinarian	Experienced	✓	✓	✓	✓
R3	Bovine Veterinarian	Experienced	✓			
R4	Vet. student	Novice	✓			
R5	Vet. student	Novice	✓	✓		
R6	Vet. student	Moderately Experienced	✓	✓		
R7	Vet. student	Novice		✓		
R8	Bovine Veterinarian	Experienced			✓	✓
R9	Bovine Veterinarian	Expert(Registered mobility scorer)			✓	✓
R10	Bovine Veterinarian	Experienced			✓	✓
R11	Bovine Veterinarian	Experienced			✓	✓
R12	Bovine Hoof trimmer	Expert(NACFT member, Category 1 qualifications)				✓
R13	Researcher	Expert (Registered mobility scorer)			✓	✓

All analyses were conducted in R (version 4.3.2), using the packages 'irr', 'stats' and 'lme4' (Bates *et al.*, 2015; Gamer *et al.*, 2019; R Core Team, 2024).

Results

Inter-assessor variation in live scoring

Inter-assessor agreement was examined for live (on-site) scoring sessions at Farm A during Visit 1 (49 cows) and Visit 2 (52 cows), using both the original 4-point AHDB mobility scale and a post-hoc dichotomised version (sound: 0–1, lame: 2–3). During Visit 1, six assessors (R1–R6) scored each cow individually. Pairwise weighted kappa values ranged from 0.12 to 0.58 (mean = 0.32, SD = 0.15; bootstrap 95% CI: 0.03–0.79), indicating slight to moderate agreement (Fig. 1). Most discrepancies occurred between scores 0 and 1, suggesting inconsistent interpretation of mild mobility changes. Agreement improved markedly when scores were dichotomised (weighted kappa range 0.052–0.86, mean = 0.45, SD = 0.21, 95% CI: –0.14 to 1.00), reflecting a simplification that removed much of the ambiguity surrounding mild mobility problems. Apparent prevalence (the proportion of cows scored as lame by each assessor – scores 2 and 3) in Visit 1 ranged from 6.1% to 22.4% among assessors (mean = 16.0% ± 5.5%, 95% CI: 10.2–21.8%).

During visit 2 five assessors mobility scored cows independently under the same conditions. Pairwise weighted kappa values ranged from 0.27 to 0.62 (mean = 0.43, SD = 0.18; 95% CI: 0.09–0.87), reflecting fair to substantial agreement (Fig. 1). As in Visit 1, disagreements were most pronounced between scores 0 and 1, reinforcing the difficulty in consistently identifying subtle or early-stage lameness. Dichotomising the scores led to a slight increase in agreement (weighted kappa ranged from 0.34 to 0.68, mean = 0.50, SD = 0.11; 95% CI: 0.05–0.85), with a higher proportion of exact matches on all assessor pairs. Apparent prevalence of lameness in Visit 2 ranged from 17.0% to 30.2% among assessors (mean = 26.4% ± 5.3%, 95% CI: 19.8–33.0%).

https://doi.org/10.1017/S0022029926102180 Published online by Cambridge University Press

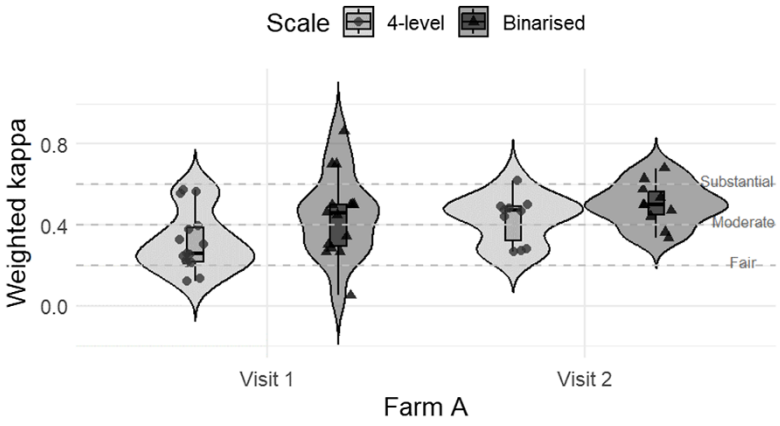


Figure 1. Distribution of inter-assessor agreement (weighted kappa values) during live mobility scoring sessions on the two Farm A visits using the original four-level AHDB scale and a dichotomised version (sound: 0–1, lame: 2–3). Each violin plot shows the distribution of pairwise kappa values between assessors for Visit 1 ($n = 15$ pairs) and Visit 2 ($n = 10$ pairs). Boxplots within violins indicate the median and interquartile range, while individual data points are overlaid. Dashed lines represent kappa interpretation thresholds according to Landis & Koch (Fair: 0.21, Moderate: 0.41, Substantial: 0.61).

Intra-assessor consistency between video and live scoring in farm a

The weighted Cohen's kappa for the first assessor was 0.332 (95% CI: 0.046–0.573), indicating fair agreement between video and live scoring. The percentage of exact matches was 49%. Similarly, the second assessor's video and live scores showed a weighted Cohen's kappa of 0.334 (95% CI: 0.010–0.574), also reflecting fair agreement, with a percentage agreement of 46.94%. These results suggest moderate intra-assessor consistency in different scoring modalities with just under half of the scores being different.

When scores were dichotomised (sound: 0–1, lame: 2–3), agreement between video and live assessments improved. For the first assessor, the weighted kappa increased to 0.557 (95% CI: 0.153–0.874), with exact matches in 89.8% of cases. For the second assessor, dichotomisation resulted in a weighted kappa of 0.264 (95% CI: 0.009–0.511) and a percentage agreement of 69.39%.

Inter-assessor variation in video scoring

Inter-assessor agreement for Farm A's video scoring was based on scores from seven independent assessors on 49 cows, using the AHDB mobility scoring system. Pairwise weighted kappa values ranged from 0.13 to 0.55 (mean = 0.31, SD = 0.12, 95% CI: 0.06–0.83), indicating slight to moderate agreement (Fig. 2). After the dichotomisation of scores, weighted kappa ranged from −0.027 to 0.68 (mean = 0.35, SD = 0.18, 95% CI: – 0.03–0.68), indicating a wider range of agreement from poor to substantial. Disagreements were found in all levels of the scores except score 3 as the dataset was not balanced and therefore there was not an equal number of cows with severely impaired mobility.

On Farm B eight assessors scored 69 cows via video using the AHDB mobility scoring system. Pairwise-weighted kappa values ranged from 0.31 to 0.71(mean = 0.50, SD = 0.1, 95% CIs ranged from 0.27 to 0.90), indicating fair to substantial agreement. After dichotomisation the agreement remained between fair and substantial, and the weighted kappa values ranged from 0.28 to 0.79 (mean = 0.53, SD = 0.12, 95% CIs: 0.05–0.87).

Assessor experience vs. agreement levels

We examined whether assessor experience was associated with differences in agreement levels in all datasets. For both Farm A onsite visits, no differences in the distribution of scores were found between the different experience groups (Visit 1: Kruskal–Wallis $\chi^2 = 6.55$, $P = 0.162$; Visit 2: Kruskal–Wallis $\chi^2 = 2.74$, $P = 0.602$). In the Farm A video assessment, the distribution of scores differed slightly (Kruskal–Wallis $\chi^2 = 7.04$, $P = 0.030$), although the magnitude of this difference was similar to that observed in Visit 1. We did not apply a multiple comparison correction due to the small number of pre-planned tests; no differences were statistically significant in the Farm B video assessments (Kruskal–Wallis $\chi^2 = 2.49$, $P = 0.288$), though expert assessors tended to show higher agreement.

Confidence level and number of views vs. score agreements

A generalized linear mixed-effects model was fitted to assess factors influencing pairwise agreement between assessors using the original AHDB scores. The model included average assessor certainty (mean of the two assessors' binary certainty scores for the same cow/video pair) and average number of views per video as fixed effects, with assessor included as a random effect to account for individual-level variation. Results for Farm A indicated a significant negative association between the average number of views and agreement probability (odds ratio = 0.61, 95% CI: 0.42–0.87, $P = 0.005$), indicating that each additional view was associated with a 39% reduction in the odds of agreement. Average certainty was negatively associated with agreement, although this effect did not reach statistical significance (odds ratio = 0.61, 95% CI: 0.35–1.05, $P = 0.08$). The model included random intercepts for assessors to account for individual differences. These results suggest that higher certainty may be associated with lower agreement, but the evidence is inconclusive.

Associations between assessor comments and mobility scores

To investigate whether comments on the video scoring were associated with mobility scores, we fitted a proportional odds logistic regression model using thematic comment categories as predictors (Fig. 3). The outcome variable was the ordinal AHDB mobility score. The proportional odds assumption was assessed using the Brant test, which did not indicate any violation. However, several combinations of comment categories and mobility scores were not observed in the dataset, which may reduce the reliability of the test. From all the comment themes, only comments related to Speed Issues showed a statistically significant association with the mobility scores (OR = 0.15, 95% CI: 0.04–0.59, $P = 0.007$), which indicates that when assessors commented on speed (e.g., 'cow moving too fast/running'), they were more likely to assign lower lameness scores. No other comment category was significantly associated

https://doi.org/10.1017/S0022029926102180 Published online by Cambridge University Press

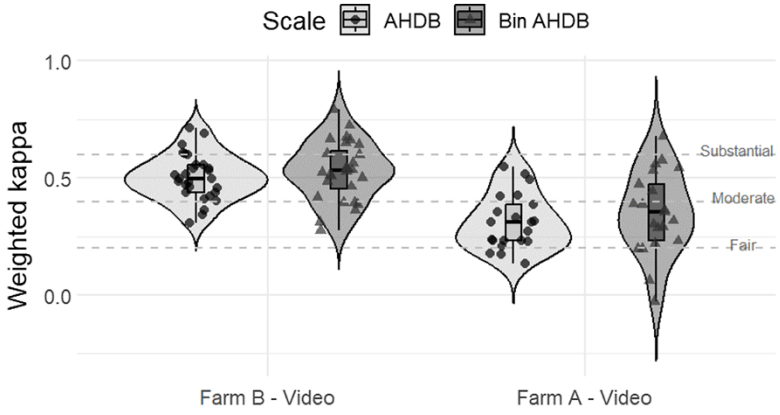


Figure 2. Distribution of weighted kappa agreement scores for assessor pairwise comparisons in Farm A and Farm B video scoring. Violin plots illustrate the density of agreement values, with overlaid boxplots showing median and interquartile ranges, and jittered points representing individual pairwise scores. Colours differentiate the scoring scales (AHDB and dichotomised). Horizontal dashed lines indicate agreement strength thresholds according to Landis & Koch interpretation (Fair = 0.21, Moderate = 0.41, Substantial = 0.61).

	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Theme 6
Theme 1	17	0	4	2	1	3
Theme 2	0	16	7	4	0	2
Theme 3	4	7	59	11	9	7
Theme 4	2	4	11	15	3	4
Theme 5	1	0	9	3	11	2
Theme 6	3	2	7	4	2	12

Theme 1 = Visibility/Video Issues
Theme 2 = Speed Issues
Theme 3 = Posture & Gait
Theme 4 = Uncertainty in Diagnosis
Theme 5 = External Physical Condition
Theme 6 = Other Scoring Notes

Figure 3. Heat-map of pairwise co-occurrence of assessor comment themes (N = 77 comments). Each cell shows the number of comments in which the two themes were mentioned together; darker colours indicate more frequent co-occurrence. Diagonal cells (theme with itself) give the total frequency of that theme in the dataset.

with the scores: Posture and Gait Observations (OR = 1.27, 95% CI: 0.45–3.61, P = 0.70), Visibility/Video Issues (OR = 0.47, 95% CI: 0.13–1.70, P = 0.28), Uncertainty in Diagnosis (OR = 0.76, 95% CI: 0.30–1.92, P = 0.65), External Physical Condition (OR = 0.50, 95% CI: 0.15–1.66, P = 0.28), and Other Technical or Personal Notes (OR = 0.52, 95% CI: 0.16–1.68, P = 0.30). The confidence intervals of all themes are wide and include values corresponding to both increases and decreases in odds of higher mobility scores, which means that the data provide limited guidance on the likely effect of these comment types.

Discussion

This study highlights the substantial subjectivity and variability inherent in human visual mobility scoring in dairy cattle. The relatively moderate inter-assessor agreement in this situation, where various individuals employed the same system of scores, illustrates the noise introduced by the human factors such as individual interpretation, experience, and surrounding environment. Although the most commonly used system in the UK (AHDB and post hoc modifications) was used, the assessors varied in their consistency of classifying cows' mobility, as reflected by the moderate weighted kappa values. This inconsistency is undesirable for research that relies on human-generated labels as ground truth for training ML algorithms. While subjectivity is a basic issue in human scoring, it also underscores the need for improvement of training methods and standardised systems to reduce variance and improve the robustness of manual assessment. The inconsistency in mobility scoring also points to the importance of identifying more objective approaches to assess animal welfare.

We found no significant differences in scores from video assessments compared to live evaluations. While this suggests that video scoring can be as reliable as live scoring, provided that the videos are recorded under conditions comparable to live observation such as similar perspective, visibility, and environmental context, the data do not rule out the possibility of meaningful differences. Effect sizes (i.e., odds ratios or log odds) would provide a more meaningful measure than P-values alone, but the sample size and score distribution in this study limited the precision of such estimates. Previous studies showed that assessors are the primary factor influencing mobility scoring outcomes, and that video-based scoring, despite potential limitations in video quality or viewpoint, is an acceptable method (Schlageter-Tello *et al.*, 2015). In this study, scores and agreement metrics were broadly similar between live and video assessments, which could be an indication that the scoring system and the human components (i.e., assessor experience and training) play potentially a more important role in the agreement than the method of assessment itself (live vs. video). However, the wide range of plausible effects indicated by confidence intervals for video-related predictors emphasises that conclusions about equivalence should be made cautiously. With proper video calibration and assessor training, video-based scoring can be a feasible alternative to live scoring in future studies, but further research using larger datasets and analyses capable of assessing both statistical and biological equivalence is required before firm conclusions can be drawn.

Although the assessors in our study had mostly strong confidence in their mobility scores, their overall agreement was low. This discrepancy shows the main problem with visual mobility scoring; confidence does not necessarily translate to accuracy or consistency. Even experienced assessors had differences between their scores, further testifying to the complexity of mobility scoring in dairy cattle. While training and calibration are commonly recommended to improve reliability (March *et al.*, 2007), our data indicate that such measures alone may not fully overcome variability, especially when applied in larger groups of assessors. Previous studies have demonstrated that experience plays no significant role in the outcome (Bokkers *et al.*, 2012; Garcia *et al.*, 2015). These results suggest that confidence in assessors' scoring decisions might be more a function of their task experience than the quality

https://doi.org/10.1017/S0022029926102180 Published online by Cambridge University Press

or consistency of their ratings, and therefore, the self-assessed confidence should not be the sole indicator of scoring reliability when training programs are being developed, or machine learning algorithms are being set for automated systems.

The variation in the human-generated labels (mobility scores) makes building robust automated systems challenging. Human-labelled data are typically used as the ground truth when training algorithms, but as the work of this paper shows, these labels are not always reliable, even for experts. Because machine learning algorithms are only as good as the data used to train them (Bakaev and Khvorostov, 2023), the variation in visual scoring can undermine the development of effective, reliable algorithms for detecting mobility issues in cows. To resolve this issue, we need to make improvements to the credibility of mobility scoring and enable the development of automated systems. We need standardised scoring protocols, which must be well-defined, explicit and exhaustive to ensure consistency across different assessors and sessions.

However, these recommendations might not be enough. Human scoring continues to be an important means of detecting mobility impairments, but the variability inherent in the method means that it cannot be the sole basis for detection, particularly of early mobility impairment. Alternative non-invasive technologies that can offer more objective, continuous monitoring of cow mobility should be investigated in future studies. Combining these technologies with full physical examinations of the animals in a hybrid system could potentially enhance early detection of mobility problems.

Conclusion

This study highlights the variation in visual mobility scoring among and between human assessors. Despite high self-reported confidence, inter-assessor agreement was modest, reflecting the subjectivity and variability in current scoring methods. Without standardised protocols, visual scores are not robust enough to serve as a solid ground truth for machine learning algorithms. The findings challenge conventional use of human-derived labels during the training of automatic lameness identification systems. In the interests of optimising accuracy and dependability, the next steps would have to outline substitute ground truth methodologies such as consensus scores, objective sensor-derived scores, or longitudinal clinical outcome evaluation.

Author contributions. K.L.: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualisation, Writing – original draft, Writing – review & editing. N.J.: Conceptualisation, Investigation, Methodology, Supervision, Visualisation, Writing – review & editing. L.V.: Investigation, Methodology, Supervision. G.K.: Investigation. J.K.: Supervision. F.F.: Supervision. Farmers, farm staff, and participating scorers: Investigation, Data curation, Resources.

Competing interests. K.L., L.V., and J.K. are involved in a prospective University of Glasgow spinout related to mobility monitoring. This work is independent of that effort. The authors declare no other competing interests.

References

Agriculture and Horticulture Development Board (2025) Mobility scoring: how to score your cows. Available at https://ahdb.org.uk/knowledge-library/mobility-scoring-how-to-score-your-cows (accessed 16 May 2025).

Amory JR, Barker ZE, Wright JL, Mason SA, Blowey RW and Green LE (2008) Associations between sole ulcer, white line disease and digital dermatitis and the milk yield of 1824 dairy cows on 30 dairy cow farms in England and Wales from February 2003-November 2004. *Preventive Veterinary Medicine* **83**(3-4), 381–391. https://doi.org/10.1016/j.prevetmed.2007.09.007

Bakaev M and Khvorostov V (2023) Quality of labeled data in machine learning: common sense and the controversial effect for user behavior models. *Engineering Proceedings* **33**(1), 3.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bokkers EAM, de Vries M, Antonissen I and de Boer IJM (2012) Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. *Animal Welfare* **21**(3), 307–318. https://doi.org/10.7120/09627286.21.3.307

Busin V, Viora L, King G, Tomlinson M, LeKernec J, Jonsson N and Fioranelli F (2019) Evaluation of lameness detection using radar sensing in ruminants. *Veterinary Record* **185**(18), 572. https://doi.org/10.1136/vr.105407

Engel B, Bruin G, Andre G and Buist W (2003) Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows. *The Journal of Agricultural Science* **140**(3), 317–333. https://doi.org/10.1017/S0021859603003983

Gamer M, Lemon J, Fellows I and Singh P (2019) *Irr: Various Coefficients of Interrater Reliability and Agreement (Version 0.84.1)*. Vienna, Austria: R Foundation for Statistical Computing. Available from Comprehensive R Archive Network (CRAN).

Garcia E, König K, Allesen-Holm BH, Klaas IC, Amigo JM, Bro R and Enevoldsen C (2015) Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. *Journal of Dairy Science* **98**(7), 4560–4571. https://doi.org/10.3168/jds.2014-9266

Green LE, Hedges VJ, Schukken YH, Blowey RW and Packington AJ (2002) The impact of clinical lameness on the milk yield of dairy cows. *Journal of Dairy Science* **85**(9), 2250–2256. https://doi.org/10.3168/jds.S0022-0302(02)74304-X

Huxley JN (2013) Impact of lameness and claw lesions in cows on health and production. *Livestock Science* **156**(1), 64–70. https://doi.org/10.1016/j.livsci.2013.06.012

Kahneman D, Rosenfield AM, Gandhi L and Blaser T (2016) Noise: how to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* **94**(10), 38–46.

March S, Brinkmann J and Winkler C (2007) Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. *Animal Welfare* **16**(2), 131–133. https://doi.org/10.1017/S096272860003116X

Mostert PF, Van Middelaar CE, De Boer IJM and Bokkers EAM (2018) The impact of foot lesions in dairy cows on greenhouse gas emissions of milk production. *Agricultural Systems* **167**, 206–212. https://doi.org/10.1016/j.agsy.2018.09.006

Nektarios S, Joseph MN, Robert FS and Georgios O (2024) Automated dairy cattle lameness detection utilizing the power of artificial intelligence; current status quo and future research opportunities. *The Veterinary Journal* **304**, 106091. https://doi.org/10.1016/j.tvjl.2024.106091

Nielsen BH, Angelucci A, Scalvenzi A, Forkman B, Fusi F, Tuyttens F, Houe H, Blokhuis H, Tind Sørensen J, Rothmann J, Matthews L, Mounier L, Bertocchi L, Marie-Madeleine R, Donati M, Per Peetz N, Salini R, de Graaf S, Hild S, Messori S, Nielsen SS, Lorenzi V, Boivin X and Thomsen PT (2014) Use of animal based measures for the assessment of dairy cow welfare ANIBAM. *EFSA Supporting Publications* **11**(9), 659E. https://doi.org/10.2903/sp.efsa.2014.EN-659

Omontese BO, Bellet-Elias R, Molinero A, Catandi GD, Casagrande R, Rodriguez Z, Bisinotto RS and Cramer G (2020) Association between hoof lesions and fertility in lactating Jersey cows. *Journal of Dairy Science* **103**(4), 3401–3413. https://doi.org/10.3168/jds.2019-17252

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schlageter-Tello A, Bokkers EA, Koerkamp PW, Van Hertem T, Viazzi S, Romanini CE, Halachmi I, Bahr C, Berckmans D and Lokhorst K (2014) Manual and automatic locomotion scoring systems in dairy cows: a review. *Preventive Veterinary Medicine* **116**(1-2), 12–25. https://doi.org/10.1016/j.prevetmed.2014.06.006

Schlageter-Tello A, Bokkers EAM, Groot Koerkamp PWG, Van Hertem T, Viazzi S, Romanini CEB, Halachmi I, Bahr C, Berckmans D and Lokhorst K (2015) Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. *Animal Welfare* **24**(1), 69–79. https://doi.org/10.7120/09627286.24.1.069

Shearer JK, Stock ML, Van Amstel SR and Coetzee JF (2013) Assessment and management of pain associated with lameness in cattle. *Veterinary Clinics: Food Animal Practice* **29**(1), 135–156. https://doi.org/10.1016/j.cvfa.2012.11.012

UK Public General Acts (1986) Animals (Scientific Procedures) Act 1986. Elizabeth II, Chapter 14. London: Her Majesty's Stationery Office (HMSO).

https://doi.org/10.1017/S0022029926102180 Published online by Cambridge University Press