The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2012, 21: 379-388
ISSN 0962-7286
doi: 10.7120/09627286.21.3.379

A training programme to ensure high repeatability of injury scoring of dairy cows

J Gibbons*†, E Vasseur‡, J Rushen† and AM de Passillé†

† Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada, PO Box 1000, 6947 Highway 7, Agassiz, BC, Canada, V0M 1A0
‡ Organic Dairy Research Centre, Universitie de Guelph, Campus d'Alfred, 31 St Paul Street, PO Box 580, Alfred, Ontario, Canada K0B 1A0
* Contact for correspondence and requests for reprints: jenny.gibbons@hotmail.co.uk

Abstract

Obtaining reliable welfare outcome measures from commercial farms can be challenging. We developed a training programme to train observers to score injuries of the tarsal joint, carpal joint and neck on dairy cows as part of an on-farm study. Twelve trainees were trained using protocols and photographs in a classroom session and on-farm visits. Continued repeatability checking was carried out during a refresher and mid-way assessment. Two trainers were used as the reference standard to which all trainees were compared. The study demonstrated that methods of scoring tarsal joint, carpal joint and neck injury can be learned by trainees from different backgrounds and high repeatability can be achieved and maintained at a very large regional or national level. Successful learning of injury scoring is dependent on protocols with strong definitions and photographs as well as repetitive training sessions. Additionally, continued repeatability checks are essential to ensure the reference standard continues to be met. This training programme can be used as a model to successfully train on-farm assessors.

Keywords: *animal welfare, dairy cattle, injury, inter-observer reliability, training, welfare assessment*

Introduction

Animal welfare assessment programmes have been designed to address public concern regarding animal welfare and there is interest in developing standardised animal welfare indicators. Measures of injuries allow us to compare the welfare of dairy cattle kept in different farming systems (eg tie-stall, free-stall and automatic milking systems) reflecting part of the welfare status of both the herd and the individual. Tarsal joint (hock), carpal joint and neck injury on dairy cows are important indicators of poor management, stall and feed-bunk design. Tarsal joint injuries are more prevalent in conventional than in organic systems (Rutherford *et al* 2008), in tie-stalls than in free-stall systems (Busato *et al* 2000), on farms with mattresses than those with sand-based free-stalls (Weary & Taszkun, 2000) and free-stalls with restricted lunge space (Potterton *et al* 2011). Additionally, tarsal joint injury is associated with increased lameness (Regula *et al* 2004; Sogstad *et al* 2005). Carpal joint lesions (Rushen *et al* 2007; Kielland *et al* 2009) are less common when cows are kept on rubber mats or mattresses in comparison to concrete-based stalls, and are more common on farms containing stalls with reduced lunge space (Haskell *et al* 2006). Cows in stalls that are frequently bedded have fewer injuries (Fulwider *et al* 2007). Presence of electric trainers and low ties rails in tie-stalls are associated with increased open tarsal joint and neck injuries, respectively (Zurbrigg *et al* 2005).

To minimise the subjectivity of welfare outcome assessment, good inter-observer agreement is paramount. Despite this, some published studies do not report inter- or intra-observer repeatability for injury scores in dairy cattle (Lombard *et al* 2010; Potterton *et al* 2011). Generally, agreement between observers is moderate to high for tarsal joint injury (Zurbrigg *et al* 2005; Rutherford *et al* 2008) but, to our knowledge, there is no published literature on observer agreement for carpal joint injury.

More recently, there is a greater emphasis on the importance of formal training programmes for animal welfare assessors to reduce inter- and intra-observer variation of animal-based measures and to maintain the integrity of the assessment (EFSA 2011; Rushen *et al* 2011). Additionally, it is important that injury scoring scales are standardised and researchers use scales that are already available in the literature (Table 1). Good training is particularly important when the assessment involves multiple observers who may not be in direct contact with each other, and who may have very different levels of experience working with animals. When training future assessors for welfare assessment, variability between people is expected due to observer-related influences such as experience and personal biases. However,

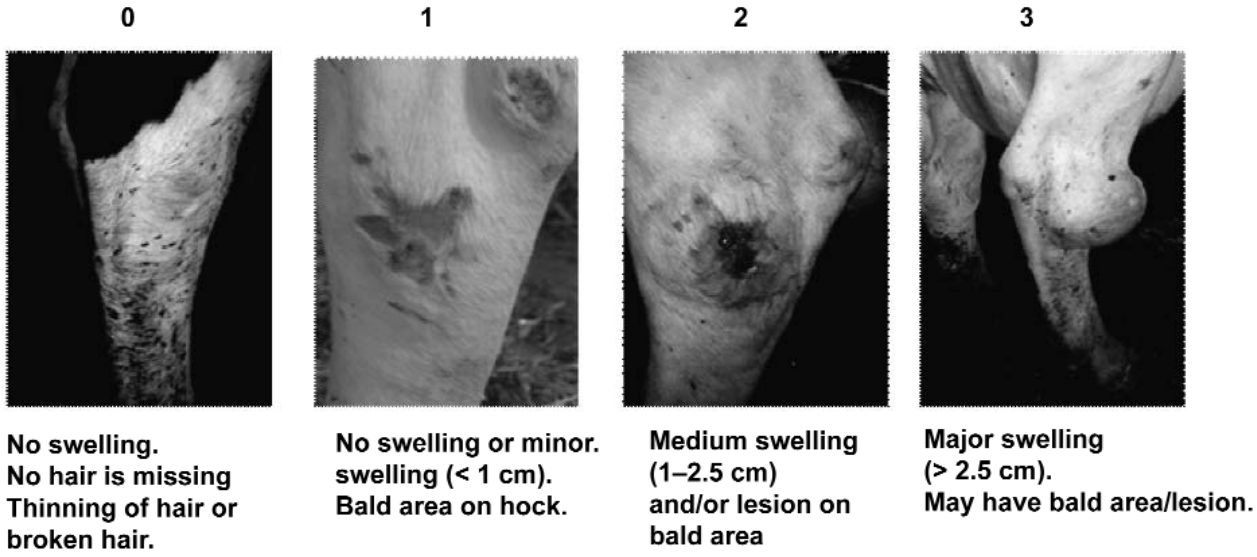
Table 1 A summary of injury scoring scales and associated cow number, observer number, observer agreement and prevalence from the following eight on-farm studies.

Reference	Stall type	Injury score	Joint	Cows (n)	Obs (n)	Inter- /intra-agreement[e]	Injury score prevalence			
							(0)	(1)	(2)	(3)
Weary & Taszkun (2000)	FS	(0) No lesion (1) Hair loss > 10cm^2, no skin breakage (2) Broken skin, dark scab, hair loss > 10cm^2	Tarsal joint	1,752	–	–	–	–	78.1	
Busato *et al* (2000)	FS TS	(0) No lesion (1) Hairloss spots (2) Superficial lesions (3) Open wounds	Tarsal joint, Carpus Knee[a]	1,886	–	–	89.6	7.3	1.6	1.5
Zurbrigg *et al* (2005)	TS	(0) No hair loss, broken skin/scabs (1) Swollen tarsal joint, no hairloss or broken skin/scab (2) Hairloss with/without swelling (3) Broken skin scabs, with/without swelling	Tarsal joint	17,894	15	80%	42.3	14.2	35.5	8.0
Zurbrigg *et al* (2005)	TS	(0) No hairloss, broken skin/scab (1) Hairloss, broken skin or scabs	Neck	17,893	15	80%	96.2	3.8		
Barberg *et al* (2007)	FS	(0) No lesion (1) Hairloss (2) Swollen tarsal joints	Tarsal joint	796	–	–	74.9	24.1	1.0	
Rutherford *et al* (2008)[b]	FS	(0) Sound: no hair damage (1) Damaged: bare patches or abrasion	Tarsal joint	–	–	84 (± 5)%	50.9	49.1		
Kielland *et al* (2009)[c]	FS	(0) No skin change (1) Hairless (2) Swollen (3) Wound	Carpal joint	2,335	5	–	65	29	5	1
Lombard *et al* (2010)	FS	(0) No hairloss/lesions (1) Hairloss/no swelling (2) Hairloss/swelling or lesion	Tarsal joint	24,825	140	–	77	20	3	
Potterton *et al* (2011)[d]	FS	(0) No hair loss (1) Mild hair loss < 2cm (2) Medium hair loss 2–2.5 cm (3) Severe hair loss > 2.5 cm	Tarsal joint	5,652	1	–	12.6	47.3	25.6	14.5
Potterton *et al* (2011)	FS	(0) No swelling (1) Mild swollen (thicker than normal) (2) Medium swollen (obviously) (3) Severely swollen (extensively)	Tarsal joint	5,877	1	–	0	74.7	23	2.3
Potterton *et al* (2011)	FS	(0) No ulceration (1) Mild ulcerated < 2 cm (2) Medium ulcerated 2–2.5 cm (3) Severely ulcerated > 2.5 cm	Tarsal joint	5,652	1	–	81.9	8.9	6.7	2.5

– Data not available in the journal article. FS = Free-Stall; TS = Tie-Stall; [a] Prevalence reported for tarsal and carpal joints combined; [b] Based on 40 non-organic farms; [c] Scoring scale used by Regula *et al* (2004); [d] Scoring scale used by Whay *et al* (2003); [e] Percentage agreement. The following studies were not included: Wechsler *et al* (2000) and Haskell *et al* (2006). They report mean number of injuries per cow.

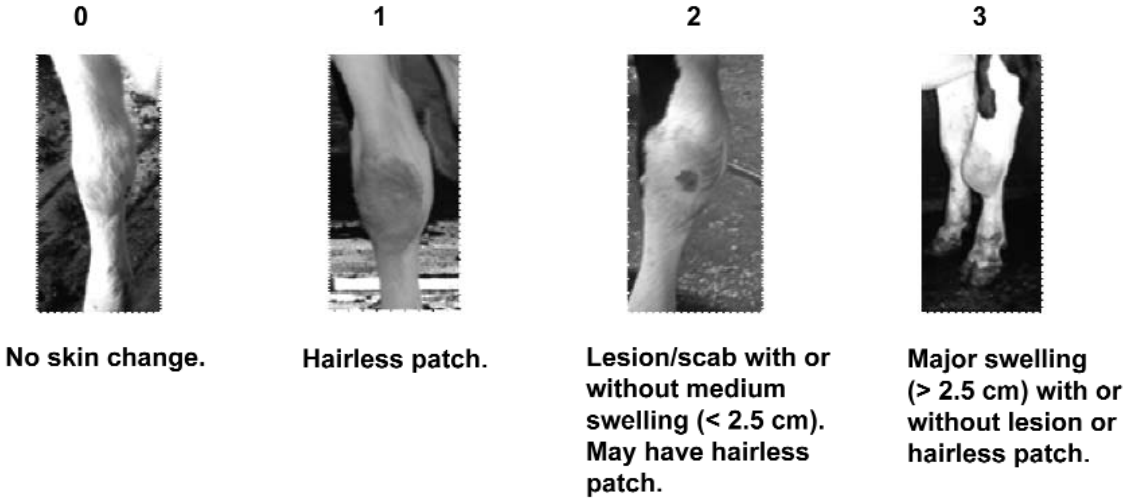
https://doi.org/10.7120/09627286.21.3.379 Published online by Cambridge University Press

Figure 1



Scoring scale for tarsal joint injury in dairy cattle.

Figure 2

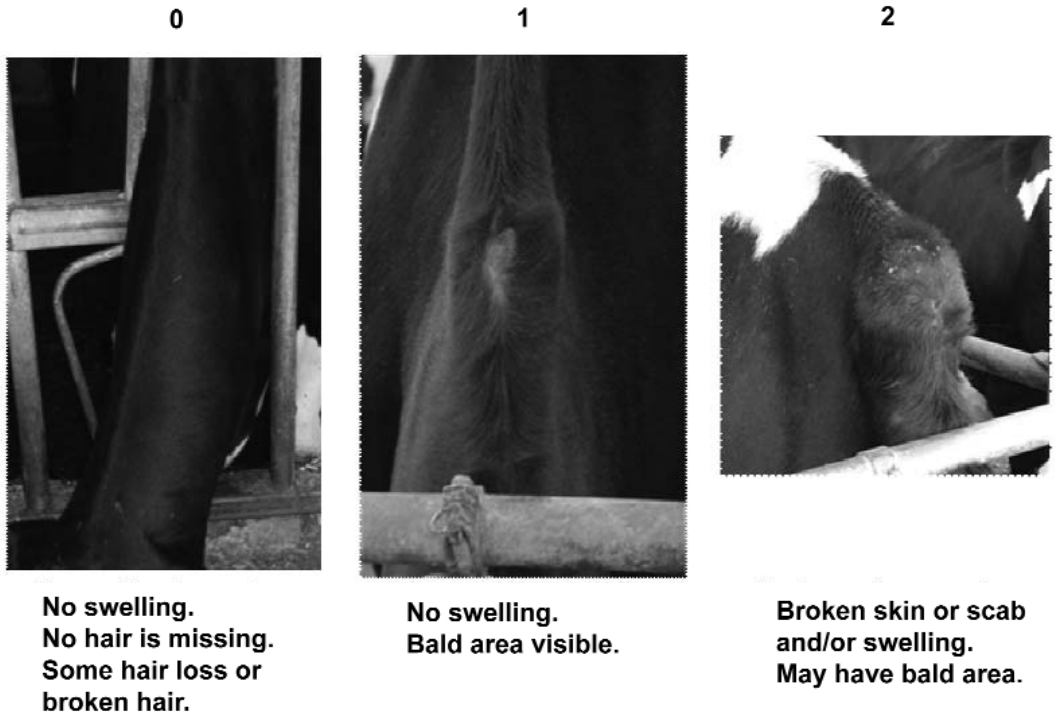


Scoring scale for carpal joint injury in dairy cattle.

with appropriate training and regular inter- and intra-observer assessment the variability in the data collected should be substantially reduced. Ideally, if different trainees receive a high standard of training with assessments at regular intervals, they should produce more accurate and reliable data (eg Mullan *et al* 2011). Despite the recognition that training is the essential component to reducing variation among observers, few studies provide detailed information on the training programme used or the effectiveness of that training.

As part of a country-wide epidemiological study on cow comfort in dairy cattle, we developed a training programme to train assessors who were naïve to the injury-scoring system, differed in previous experience with dairy cattle, and who were geographically separated, with little direct contact. This paper reports the use of a training programme and regular assessment to achieve high repeatability of scores and therefore more reliable data.

Figure 3



Scoring scale for neck injury in dairy cattle.

Materials and methods

Development of protocol

Tarsal joint, carpal joint and neck injury scoring protocols were developed by the trainers. For the initial development of the protocols, six observers systematically tested different sampling methods such as number of observers required, position of observer in relation to the tarsal joint, carpal joint or neck and scoring location (eg headlock, milking parlour, inside the pen or outside the pen). The different sampling methods were analysed for strengths and weaknesses and the outcome of this assisted to improve the practicality of assessing injuries. The improved tarsal joint, carpal joint and neck injury protocols were distributed to a group of dairy experts for further improvement to refine description of each injury score. The finalised protocols contained definitions of each injury score along with a representative photograph placed on a laminated reference card, and a detailed description of the procedures for taking the measures. This later outlined factors such as the distance to stand from the cows, the angle of observation, etc. In addition to the protocol, a summary table of the scoring system with concise definitions was placed on the data recording sheet.

Injury scores

A tarsal joint injury scoring system was adapted from the Cornell University Cooperative Extension Hock Assessment for Cattle (http://www.ansci.cornell.edu/prodairy/pdf/hockscore.pdf) (Figure 1). The condition of the lateral surface of the left and right tarsal joints, not including the tuberosity of calcaneous (point of the hock), was recorded using a 0–3 scale: No swelling with minor or no hair loss or broken hairs (0); No swelling or minor swelling with thickness of < 1 cm with bald area (1); Medium swelling thickness of 1–2.5 cm and/or lesion/scab, may have bald area (2); Major swelling thickness of > 2.5 cm (3). The condition of the anterior surface of the left and right carpal joints was recorded using a four-point scale: No swelling with minor or no hair loss or broken hairs (0); No swelling with bald area (1); Swelling with thickness of < 2.5 cm and/or broken skin or scab, may have bald area (2); Major swelling with thickness of ≥ 2.5 cm, with or without bald area or lesion (3) (Figure 2). In tie-stalls, the left and right tarsal joints were scored by an observer standing at the back of the stall and carpal joint injury was scored from the front as the cow stood stationary in her stall. In free-stalls, tarsal joint injury was scored from the side in the milking parlour and carpal joint injury was scored from the front either at a head-locked feed-bunk or as the cow stood in a stall or alley of the pen. Tarsal and carpal joints were both examined from a maximum distance of 50 cm. Tarsal and carpal joints were not scored if they were too dirty or not visible (eg herringbone parlour).

A neck injury scoring system was adapted from the literature (Zurbrigg *et al* 2005; Lapointe 2010) (Figure 3). The

https://doi.org/10.7120/09627286.21.3.379 Published online by Cambridge University Press

condition of the dorsal surface of the neck from directly behind the ears up to the point directly above the shoulder joint was recorded using a three-point scale: No swelling with minor or no hair loss or broken hairs (0); Bald area with no swelling (1) and; Broken skin or scab and/or swelling with or without bald area (2). Recording was carried out on each cow standing in the tie-stall and in free-stalls at a head-locked feed-bunk. Cows were examined from the front and always from a maximum distance of 50 cm.

Implementation of training programme

Photographic training aids

Digital, coloured photographs were used to demonstrate each injury score on the laminated reference cards and during training in a Powerpoint presentation. The photographs were always taken within 50 cm of the animal and demonstrated tarsal joints, carpal joints and necks ranging from healthy to injured.

Trainers

The two trainers were both experienced dairy scientists with extensive experience of scoring cow injuries on commercial farms. Using the finalised protocols, the trainers scored and discussed injury scores on cows until they developed uniform scoring. In addition, they underwent four scoring repeatability sessions including two on-farm and two photograph sessions to ensure a high level of agreement. To maintain agreement, inter-repeatability was assessed monthly during the trainee training period (six-month period) by scoring 20 cows live in the barn or remotely using photographs. The trainers set the reference standard against which each trainee was evaluated throughout the entire training programme. Trainer A was responsible for the training of the trainees and trainer B for continued assessment during field data collection. Only lactating Holstein cows were used throughout the training programme and the cow's identity was recorded using the ear-tag number.

Training programme

There were a total of 12 trainees with varying degree of experience working with dairy cattle: six trainees had more than four-years experience working with dairy cattle and six trainees had no experience working with dairy cattle prior to commencing the training programme. The trainees consisted of three teams with a team located in Ontario, Quebec and Alberta in Canada. The trainers were located in British Columbia in Canada. Each team received an identical training programme between January and June 2011. The same trainer delivered the course which started with a 2-h classroom instruction session followed by a 2-h live session on a research dairy unit.

In the classroom, a presentation outlining the rationale and protocol of the measures was given. Uncertainties about the scoring of the different measures were discussed with the help of eight photographs for each location. On day 1, the trainees were tested with photographs of tarsal joint (n = 20), carpal joint (n = 20) and neck injury (n = 20) previously scored by the two trainers. Only nine trainees scored the neck injury photographs. The agreement scores of the two trainers were used as a reference standard and compared with those scores given by the trainees. Later on day 1, during the live session, trainees scored injuries on cows (n = 20) in the research dairy unit. On day 2, the trainees scored injuries on cows (n = 20) in the research dairy unit. Six and seven days after day 1, trainees scored 20 cows each on two commercial farms (Day 7, Day 8). Only nine trainees scored on Day 8. When the target of a weighted kappa (K_w) > 0.6 with the trainer was achieved, the trainee was considered trained for that measure. Trainees not reaching this target by Day 8, did not continue to score injury on commercial farms.

The training took place on a total of three different research dairy units (two tie-stalls and one free-stall) and six different commercial farms (four tie-stall and two free-stall), all with Holstein cattle. Firstly, the trainer and the trainees both scored the same cows (n = 20) and results were compared. Secondly, the trainer and trainees discussed the measures that had low agreement during the classroom session and that were more easily trained on-farm. Finally, the trainer explained the sampling procedure, as well as practical, safety aspects to take into account when approaching animals and discussed any potential challenges that may be encountered in the field (eg to gain optimum visibility of tarsal joints, carpal joints and necks). In farms where lighting was poor, observers were advised to shine a flash lamp towards the assessment area to improve accuracy.

Refresher course and mid-way check

Once field data collection was in progress, the trainees were re-assessed twice to ensure that they remained objective, impartial and repeatable in their scoring. A refresher course carried out 3–4 weeks after initial training was completed and involved the trainees scoring Powerpoint presentations of photographs of tarsal-joint (n = 32) injury. For the mid-way assessment, which was completed between 5–15 weeks after the initial training, trainer B accompanied the trainees on-farm. The trainer and trainees scored tarsal joint (n = 80), carpal joint (n = 80) and neck injury (n = 40) on two commercial farms. Repeatability between trainer and each trainee was calculated and discrepancies were discussed.

Statistical analysis

The data were entered in Microsoft Excel (Microsoft 2007) and analysed using SAS software (version 9.2; SAS Institute 2008). Both percent agreement and weighted Kappa coefficient (K_w) will be presented to make it easier to cross-reference to other studies. Percent agreement and K_w was calculated between trainers as well as between trainers and trainees. The percent exact agreement was calculated as: number of exact agreement/total number of observations × 100. K_w statistic was used to assess the extent to which the proportion of agreement within or between observers is better than chance. In this way, K_w is more stringent than correlations or raw percentage agreement alone (Hoehler 2000). The interpretation of K_w values

Table 2 The percentage of tarsal joints, carpal joints and necks assigned each score during the training programme based on trainer reference scores.

Injury measure	Time point[1]	n[2]	Injury ordinal scale[3]			
			0	1	2	3
Tarsal joint	Day 1, Day 2	355	47.8	12.5	34.5	5.2
	Day 7, Day 8	136	57.2	13.5	28.5	0.8
	Mid-way	442	39.3	20.4	37.8	2.5
Carpal joint	Day 1, Day 2	358	61.5	9.6	28.2	0.7
	Day 7, Day 8	136	61.2	10	28.5	0.3
	Mid-way	429	54.3	8.6	35.7	1.4
Neck	Day 1, Day 2	180	58.8	19.1	22.1	–
	Day 7, Day 8	68	76.7	3.9	19.4	–
	Mid-way	230	63.9	19.6	16.5	–

[1] On day 1 of the training programme, injury was scored from photograph as well as live, on day 2, 7 and 8 injury was scored live. Injury was scored live during the midway assessment which was carried out 5–15 weeks after the training programme.
[2] n = number of tarsal joints, carpal joints and necks scored.
[3] Tarsal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling or minor swelling with thickness of < 1 cm with bald area; (2) Medium swelling thickness of 1–2.5 cm and/or lesion/scab on bald area; (3) Major swelling thickness of > 2.5 cm. May have bald area or lesion.
Carpal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling with bald area; (2) Swelling with thickness of < 2.5 cm and/or broken skin or scab, may have bald area; (3) Major swelling with thickness of ≥ 2.5 cm, with or without bald area or lesion.
Neck Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) bald area with no swelling; (2) Broken skin or scab and/or swelling with or without bald area.

according to Landis and Koch (1977) is: < 0 = poor; 0.0–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; and 0.81–1 = almost perfect. The target level of $K_w > 0.6$ needed to be reached during training and maintained at the midway check in order to ensure a high level of agreement. The discrepancy between trainer and trainee for each time-point during the training programme (Day 1 [Photograph and Live], Day 2, Day 7, Day 8, Refresher, Mid-way) was tallied for each combination of scores (0–1; 0–2; 0–3; 1–2; 1–3; 2–3) and converted to a percentage of the total. A Kruskal-Wallis non-parametric test was used to test the effect of the trainee's previous dairy experience (> four years or none) for each time point during the training programme (Day 1 [Photograph and Live], Day 2, Day 7, Day 8, Refresher, Mid-way).

In addition to testing repeatability of the multi-category ordinal scales for tarsal joint, carpal joint and neck injury scores, the ordinal scales were collapsed to form a binary scale. Tarsal joint and carpal joint scores 0 and 1 (no skin change, hairless patches) were collapsed together and contrasted with the combined score of 2 and 3 (broken skin or scabs with obvious swelling). Combined neck score 0 and 1 (no skin change, hairless patches) were contrasted with score 2 (broken skin or scab with or without swelling). This way, only cows with severe differences from the normal condition were classified as injured.

Results

Prevalence of tarsal joint, carpal joint and neck injury

The percentage of tarsal joints and carpal joints with no injury ranged across all time points from 39.3 to 61.2%, for bald area and minor swelling the range was 8.6–20.4%, for medium swelling and/or lesion 28.2–37.8% and for major swelling 0.3–5.2% (Table 2). The range for no neck injury was 58.8–76.7%, no bald area was 3.9–19.6 and no swelling was 16.5–22.1 across all time points (Table 2).

Intra- and inter-observer repeatability (for trainers)

Intra- and inter-repeatability of the trainers remained consistently high (> 70%; $K_w > 0.64$) for all measures throughout a six-month period. The mean exact agreement between the trainers for tarsal joint, carpal joint, and neck was 94, 93 and 100%, respectively. The mean K_w values between the trainers for tarsal joint, carpal joint and neck was 0.94, 0.81 and 1, respectively. The mean intra-observer exact agreement for tarsal joint, carpal joint and neck ranged from 81–97% and K_w 0.63–0.95 for trainer A and from 85–100% and K_w 0.86–1 for trainer B.

Inter-observer repeatability (trainee and trainer)

The results of all trainees are presented in this paper where possible. A few observers failed to record injury score during some time points of the programme. Table 3 reports the average value of the K_w, confidence interval

https://doi.org/10.7120/09627286.21.3.379 Published online by Cambridge University Press

Table 3 Mean number (n), weighted Kappa coefficient K_w (95% confidence interval) and percentage agreement (min-max) for tarsal joint, carpal joint and neck injury across the time points of the training programme for both the ordinal scale and the binary scoring systems.

Injury measure	Time point[1]	No observers	n[2]	Injury ordinal scale[3]		Injury binary scale[4]
				K_w [95% CI]	% [Min–Max]	% [Min–Max]
Tarsal joint	Day 1: Photo	12	240	0.70 [0.49–0.92]	69.83 [55–85]	84.58 [75–95]
	Day 1: Live	12	532	0.73 [0.55–0.90]	78.17 [65–90]	88.57 [73.91–100]
	Day 2	12	352	0.80 [0.62–0.96]	83.79 [70–100]	91.87 [76.67–100]
	Day 7	12	489	0.61 [0.41–0.80]	75.33 [50–98]	81.10 [51.28–100]
	Day 8	9	294	0.75 [0.53–0.96]	84.33 [73–92]	95.56 [84.62–100]
	Refresher*	6	32	0.70 [0.50–0.90]	72.30 [31–88]	84.38 [78.13–90.63]
	Mid-way*	6	669	0.68 [0.53–0.83]	73.33 [64–82]	82.05 [66.18–96.08]
Carpal joint	Day 1: Photo	12	476	0.37 [0.14–0.60]	56.58 [40–73]	68.28 [47.5–85]
	Day 1: Live	12	514	0.51 [0.41–0.71]	63.50 [31–96]	73.33 [43.75–100]
	Day 2	12	331	0.48 [0.22–0.75]	71.17 [54–90]	81.69 [69.23–100]
	Day 7	12	494	0.42 [0.17–0.67]	68.25 [53–93]	75.16 [55.56–100]
	Day 8	9	294	0.58 [0.34–0.81]	73.44 [64–87]	84.11 [65.38–95.65]
	Mid-way*	6	669	0.69 [0.59–1.00]	80.83 [74–100]	86.21 [69.64–112.77]
Neck	Day 1: Photo	9	137	0.60 [0.32–0.88]	73.86 [52–90]	83.06 [65–100]
	Day 1: Live	12	241	0.56 [0.26–0.86]	75.25 [40–92]	89.42 [73.91–100]
	Day 2	12	166	0.60 [0.26–0.89]	76.83 [46–100]	90.07 [73.68–100]
	Day 7	12	237	0.54 [0.22–0.81]	69.58 [39–95]	87.59 [43.48–100]
	Day 8	9	144	0.75 [0.49–0.95]	85.44 [64–100]	87.90 [66.67–100]
	Mid-way*	6	352	0.71 [0.52–0.90]	83.67 [68–96]	96.05 [66.67–100]

[1] On day 1 of the training programme, injury was scored from photograph as well as live, on day 2, 7 and 8 injury was scored live. Injury was scored live during the mid-way assessment which was carried out 5–15 weeks after the training programme.

[2] n = number of tarsal joints, carpal joints and necks scored.

[3] Tarsal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling or minor swelling with thickness of < 1 cm with bald area; (2) Medium swelling thickness of 1–2.5 cm and/or lesion/scab on bald area; (3) Major swelling thickness of > 2.5 cm. May have bald area or lesion.

Carpal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling with bald area; (2) Swelling with thickness of < 2.5 cm and/or broken skin or scab, may have bald area; (3) Major swelling with thickness of ≥ 2.5 cm, with or without bald area or lesion.

Neck Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) bald area with no swelling; (2) Broken skin or scab and/or swelling with or without bald area.

[4] Binary scale = 0 + 1 versus 2 + 3 for tarsal joints and carpal joints; 0 + 1 versus 2 for neck.

* This data only include observers that reached the K_w > 0.6 on Day 8 and continued to score for the rest of the study.

and percentage exact agreement. Percent exact agreement ranged from low to high on day 1 (Photograph and Live) for tarsal joints, carpal joints and neck (Table 3). Between Day 2 and Day 7 agreement decreased but improved again on Day 8 for tarsal joints, carpal joints and neck (Table 3). When the binary scale was used agreement was higher at all time points (Table 3). This shows that differences between trainees and the trainer were less for the binary scale than for the full ordinal scale. A total of 10, 6 and 10 trainees reached the target agreement of $K_w > 0.6$ for tarsal joint, carpal joint and neck, respectively by Day 8. There is a higher level of discrepancy for carpal joint injury score between trainer and trainees compared to tarsal joint and neck injury (Table 4). More specifically, there is a greater occurrence of discrepancy between 0–3 and 1–3 with carpal joint injury compared to tarsal joint injury (Table 4).

The trainee's previous dairy experience did not significantly affect repeatability of tarsal joint, carpal joint and neck injury at any of the time-points during the training programme (Day 1 [Photograph and Live], Day 2, Day 7, Day 8, Refresher, Mid-way) ($P > 0.05$).

Table 4 The occurrence (%) of discrepancies and agreement between trainer and trainee for each combination of tarsal joint, carpal joint and neck injury scores.

Injury measure	Time point[1]	n[2]	Injury ordinal scale[3]						
			0–1	0–2	0–3	1–2	1–3	2–3	Agree
Tarsal joint	Day 1: Photo	240	11.67	2.08	0	13.33	0	2.92	70
	Day 1: Live	532	6.02	1.88	0	9.59	0	1.32	81.20
	Day 2	352	4.26	1.99	0	6.53	0	1.99	85.23
	Day 7	489	6.75	10.02	0	8.38	0	0.82	73.62
	Day 8	294	10.20	1.36	0	2.38	0	0.68	85.37
	MW	667	7.65	4.80	0	11.99	0	1.95	73.61
Carpal joint	Day 1: Photo	476	8.40	11.34	0.63	19.54	0.21	2.94	56.93
	Day 1: Live	514	7.59	12.06	0.97	14.79	0.19	0.78	63.62
	Day 2	331	12.08	10.57	0.60	6.65	0.30	0.30	69.49
	Day 7	494	9.31	18.42	0.20	6.88	0	0	65.18
	Day 8	294	10.88	6.80	0	8.5	0	0	73.81
	MW	669	9.87	15.10	0	4.63	0	0.60	69.81
Neck	Day 1: Photo	137	6.57	6.57		10.22			76.64
	Day 1: Live	240	16.25	7.92		2.92			72.92
	Day 2	166	14.46	7.23		4.22			74.10
	Day 7	237	13.92	10.97		2.53			72.57
	Day 8	144	15.97	9.03		1.39			73.61
	MW	352	8.81	8.81		3.69			78.69

[1] On day 1 of the training programme, injury was scored from photograph as well as live, on day 2, 7 and 8 injury was scored live. Injury was scored live during the mid-way assessment which was carried out 5–15 weeks after the training programme.
[2] n = number of tarsal joints, carpal joints and necks scored.
[3] Tarsal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling or minor swelling with thickness of < 1 cm with bald area; (2) Medium swelling thickness of 1–2.5 cm and/or lesion/scab on bald area; (3) Major swelling thickness of > 2.5 cm. May have bald area or lesion.
Carpal Joint Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) No swelling with bald area; (2) Swelling with thickness of < 2.5 cm and/or broken skin or scab, may have bald area; (3) Major swelling with thickness of ≥ 2.5 cm, with or without bald area or lesion.
Neck Injury Score = (0) No swelling with minor or no hair loss or broken hairs; (1) bald area with no swelling; (2) Broken skin or scab and/or swelling with or without bald area.

Discussion

Our training programme achieved good agreement between trainer and trainees for tarsal joint injury and improved the agreement between trainer and trainees for carpal joint and neck injury. A high level of agreement was maintained during a six-month period. The training programme was equally successful in training people with and without dairy experience. This demonstrates that high repeatability of injury scoring can be achieved at a large regional or national level despite the large distances between trainees and trainers, and despite differences between trainees in their prior experience with dairy cattle.

Additionally, it was best practice to remove people with poor repeatability from scoring injuries: trainees not reaching the target of $K_w = 0.6$ on Day 8, did not continue to score injury on-farm but instead scored for a measure in which the target level was achieved (eg Body Condition Score). It is acknowledged that poor agreement may highlight that more in-depth training is required. However, sometimes additional training is not possible, so trainees that do not meet a target level of agreement should not be used for scoring. It was essential to our study that the two trainers achieve a high inter- and intra-repeatability. The use of the trainers as the reference point to assess agreement was appropriate in this study as the trainees were located in different parts of Canada. Trainers have also been used as the reference point in other studies, for example in assessing donkey and pig welfare (Pritchard *et al* 2007; Mullan *et al* 2011).

During the six-month period after training, repeatability was assessed at least twice for each trainee. As a complement to that, the trainees reviewed the protocols regularly and scored injury on a weekly basis. The target level for tarsal joints, carpal joints and necks was still achieved at the mid-way assessment (5–15 weeks after initial training) for

https://doi.org/10.7120/09627286.21.3.379 Published online by Cambridge University Press

all trainees that met the training target on Day 8. In addition to checking repeatability in person, our study highlights that repeatability can conveniently be done remotely using a Powerpoint presentation of good quality photographs as we did in the refresher assessment.

The agreement between trainee and the trainer achieved during the training programme is within the range of the sparse information reported on repeatability of tarsal joint injuries within the literature (Zurbrigg *et al* 2005; Thomsen & Baadsgaard 2006; Rutherford *et al* 2008). Our intensive training programme achieved a minimum repeatability of 73% and $K_w = 0.53$ for a four-point ordinal tarsal joint injury scale by Day 8. Repeatability of carpal joint injury scores was constantly lower than tarsal joint and neck injury scores. However, repeatability of carpal joint injury scores increased gradually over the training week but only six trainees achieved the target level of agreement by Day 8. Our carpal joint scoring system was adapted from a tarsal joint injury scale because of the lack of scoring systems available in the literature. It is possible that a more appropriate scale would be suited for carpal joint injury. Alternatively, the reduced agreement for carpal joint injury might be due to the fact that it is identical to the tarsal joint scale apart from score 1 which differs slightly. The reduced agreement may be attributed to forgetting this difference between the scales or being too over confident and recalling the scoring system incorrectly from memory.

The five-day break between Days 2 and 7 resulted in decreased agreement for all injury scores on Day 7 but the agreement improved again on Day 8. This highlights the importance of continual practice, particularly during the sensitive learning phase. It is extremely important that repeatability be continually checked at specific time points during data collection and it is not sufficient to carry out one test of repeatability at the beginning. Many countries worldwide have welfare audits being implemented at national level, and often repeatability assessment is not repeated once the assessor has been trained.

It was important that, throughout the training programme, trainees were exposed to a sufficient number of animals from each of the level of the scores. In the literature (see Table 4), tarsal joint injury prevalence ranges from 23 to 78.1% in free-stalls (Weary & Taszkun 2000; Lombard *et al* 2010) and prevalence up to 57.7% in tie-stall (Zurbrigg *et al* 2005). The tarsal joint injury prevalence during our study ranged from 42.8 to 60.7% when the prevalence of tarsal joints with hair loss, lesion and swelling were summed (score 1, 2, 3). There are few studies available on carpal joint injury, however, Kielland *et al* (2009) reported 35% prevalence of carpal joints with hair loss, swelling and lesions. This is slightly lower than the range reported in our study from 38.5 to 45.7% when the prevalence of carpal joints with hair loss, lesion and swelling are summed (score 1, 2, 3). The prevalence of neck injury presented by Busato *et al* (2000) was 1.3% and by Zurbrigg *et al* (2005) was 3.8% which is substantially less than the 23.3–41.2% prevalence of neck injury in our sample.

The majority of the errors in scoring injury were between neighbouring scores on the scale, eg 0 and 1, 1 and 2. In our study, lesion or scab size were not quantified unlike in Wechsler *et al* (2000), Whay *et al* (2003) and Potterton *et al* (2011) and this may explain the discrepancies between scores 1 and score 2. For this reason, we collapsed the scale to a binary scoring system and achieved higher agreement between the trainer and trainees compared to our four-point ordinal scale. For tarsal joints, 96% agreement on Day 8 is higher than the agreement of 85% reported by Rutherford *et al* (2008) for a similar binary tarsal joint injury score. Similarly, Thomsen and Baadsgaard (2006) reported higher agreement for observers when they collapsed an ordinal tarsal joint lesion scale to form a binary scale (ie healthy or injured). In general, the optimal choice for a scoring scale depends on the purpose of the specific study. For example, in an epidemiological study, more detail will be required at the individual level, which may require an ordinal scale, whereas during a welfare assessment audit, information will be required at the herd level, for which a binary scale may be sufficient.

During the course of this work some practical factors affecting the accuracy of injury scoring were identified. Accuracy of injury assessment is dependent on good lighting and the distance in which the cow can be approached in free-stall systems, particularly those systems in which cows cannot be head-locked or scored in the parlour. Bald patches and lesions are more easily identified on white hair with pink skin compared to black hair and black skin. During assessment of swelling, one carpal joint was compared to the other, however, carpal joints are not always symmetrical which makes it very challenging to assess swelling. It is important that neck assessment is carried out by the observers when the neck of the animal is relaxed (eg during eating). Scoring outstretched necks during eating compared to scoring relaxed necks when animals are in a head-up state may yield different assessments particularly for swellings. This may have contributed to the error in neck swellings between the experienced trainers and the trainees.

Recommendations

This study describes a training programme that successfully obtains a high level of repeatability between assessors and trainers of injury scoring of dairy cows. This training programme demonstrated the necessity to perform regular repeatability assessment and that simpler scoring scales can provide more reliable results when compared to a more precise scoring scale. Therefore, it is recommended that on-farm studies should use simple scales with a robust training programme to include regular repeatability checks.

Animal welfare implications and conclusion

Accurate assessment of animal welfare is the first step to improving animal welfare. Accredited assurance systems using standardised welfare outcome measures for dairy cattle could potentially be a useful way to deliver welfare assessment of a large number of dairy cattle. However, before this can be achieved standard procedures for training and checking need to be developed to produce reliable data and prevent bias in interpretation of results.

Acknowledgements

This study was funded by Agriculture & Agri-Food Canada and Dairy Farmers of Canada (Ottawa, Ontario, Canada) as part of the Dairy Science Cluster initiative. We thank the collaborators, students and co-op students from Agriculture & Agri-Food Canada (Agassiz, British Columbia, Canada), University of British Columbia (Vancouver, British Columbia, Canada), University of Calgary (Calgary, Alberta, Canada), University of Guelph (Guelph, Ontario, Canada), Université Laval (Quebec City, Quebec, Canada), and Valacta (Sainte-Anne-de-Bellevue, Quebec, Canada).

References

Barberg AE, Endres MI, Salfer JA and Reneau JK 2007 Performance and welfare of dairy cows in an alternative housing system in Minnesota. *Journal of Dairy Science 90*: 1575-1583. http://dx.doi.org/10.3168/jds.S0022-0302(07)71643-0

Busato A, Trachsel P and Blum JW 2000 Frequency of traumatic cow injuries in relation to housing systems in Swiss organic dairy herds. *Journal of Veterinary Medicine 47*: 221-229. http://dx.doi.org/10.1046/j.1439-0442.2000.00283.x

EFSA 2011 *Scientific opinion on the use of animal-based measures to assess the welfare of dairy cows EFSA panel on animal health and welfare.* European Food Safety Authority (EFSA): Parma, Italy

Fulwider WK, Grandin T, Garrick DJ, Engle TE, Lamm WD, Dalsted NL and Rollin BE 2007 Influence of free-stall base on tarsal joint lesions and hygiene in dairy cows. *Journal of Dairy Science 90*: 3559-3566. http://dx.doi.org/10.3168/jds.2006-793

Haskell MJ, Rennie LJ, Bowell VA, Bell MJ and Lawrence AB 2006 Housing system, milk production, and zero-grazing effects on lameness and leg injury in dairy cows. *Journal of Dairy Science 89*: 4259-4266. http://dx.doi.org/10.3168/jds.S0022-0302(06)72472-9

Hoehler FK 2000 Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology 53*: 499-503. http://dx.doi.org/10.1016/S0895-4356(99)00174-2

Kielland C, Ruud LE, Zanella AJ and Østerås O 2009 Prevalence and risk factors for skin lesions on legs of dairy cattle housed in freestalls in Norway. *Journal of Dairy Science 92*: 5487-5496. http://dx.doi.org/10.3168/jds.2009-2293

Landis JR and Koch GG 1977 The measurement of observer agreement for categorical data. *Biometrics 33*: 159-174. http://dx.doi.org/10.2307/2529310

Lapointe GD 2010 Vos Vaches Sont-Elles. *Proceedings of the 34th Symposium sur les Bovins Laitiers* pp 119-142. Québec, QC, Canada

Lombard JE, Tucker CB, von Keyserlingk MAG, Kopral CA and Weary DM 2010 Associations between cow hygiene, hock injuries, and free stall usage on US dairy farms. *Journal of Dairy Science 93*: 4668-4676

Mullan S, Edwards SA, Butterworth A, Whay HR and Main DCJ 2011 Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. *The Veterinary Journal 190*: 100-109. http://dx.doi.org/10.1016/j.tvjl.2011.01.012

Potterton SL, Green MJ, Harris J, Millar KM, Whay HR and Huxley JN 2011 Risk factors associated with hair loss, ulceration, and swelling at the hock in freestall-housed UK dairy herds. *Journal of Dairy Science 94*: 2952-2963. http://dx.doi.org/10.3168/jds.2010-4084

Pritchard JC, Barr ARS and Whay HR 2007 Repeatability of a skin tent test for dehydration in working horses and donkeys. *Animal Welfare 16*: 181-183

Regula G, Danuser J, Spycher B and Wechsler B 2004 Health and welfare of dairy cows in different husbandry systems in switzerland. *Preventive Veterinary Medicine 66*: 247-264. http://dx.doi.org/10.1016/j.prevetmed.2004.09.004

Rushen J, Butterworth A and Swanson JC 2011 Animal behavior and well-being symposium: farm animal welfare assurance: science and application. *Journal of Animal Science 89*: 1219-1228. http://dx.doi.org/10.2527/jas.2010-3589

Rushen J, Haley D and de Passillé AM 2007 Effect of softer flooring in tie stalls on resting behavior and leg injuries of lactating Cows. *Journal of Dairy Science 90*: 3647-3651. http://dx.doi.org/10.3168/jds.2006-463

Rutherford KMD, Langford FM, Jack MC, Sherwood L, Lawrence AB and Haskell MJ 2008 Hock injury prevalence and associated risk factors on organic and nonorganic dairy farms in the United Kingdom. *Journal of Dairy Science 91*: 2265-2274. http://dx.doi.org/10.3168/jds.2007-0847

SAS Institute 2008 *SAS User's Guide.* SAS Institute Inc: Cary, NC, USA

Sogstad AM, Fjeldaas T and Osteras O 2005 Lameness and claw lesions of the Norwegian red dairy cattle housed in free stalls in relation to environment, parity and stage of lactation. *Acta Veterinaria Scandinavica 46*: 203-217. http://dx.doi.org/10.1186/1751-0147-46-203

Thomsen PT and Baadsgaard NP 2006 Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. *Preventive Veterinary Medicine 75*: 133-139. http://dx.doi.org/10.1016/j.prevetmed.2006.02.004

Vapnek J and Chapman M 2010 *Legislation and regulatory options for animal welfare.* FAO: Rome, Italy

Weary DM and Taszkun I 2000 Hock lesions and free-stall design. *Journal of Dairy Science 83*: 697-702. http://dx.doi.org/10.3168/jds.S0022-0302(00)74931-9

Wechsler B, Schaub J, Friedli K and Hauser R 2000 Behaviour and leg injuries in dairy cows kept in cubicle systems with straw bedding or soft lying mats. *Applied Animal Behaviour Science 69*: 189-197. http://dx.doi.org/10.1016/S0168-1591(00)00134-9

Whay HR, Main DCJ, Green LE and Webster AJF 2003 Assessment of the welfare of dairy cattle using animal-based measurements: direct observations and investigation of farm records. *Veterinary Record 153*: 197-202. http://dx.doi.org/10.1136/vr.153.7.197

Zurbrigg K, Kelton D, Anderson N and Millman S 2005 Tie-stall design and its relationship to lameness, injury, and cleanliness on 317 Ontario dairy farms. *Journal of Dairy Science 88*: 3201-3210