J. Dairy Sci. 102:5466–5474
https://doi.org/10.3168/jds.2018-15644

Interobserver agreement of digital dermatitis M-scores for photographs of the hind feet of standing dairy cattle

A. Vanhoudt,[1]* D. A. Yang,[2] T. Armstrong,[3] J. N. Huxley,[2] R. A. Laven,[2] A. D. Manning,[4] R. F. Newsome,[5] M. Nielen,[1] T. van Werven,[1,6] and N. J. Bell[7]

[1]Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, 3584 CL, Utrecht, the Netherlands
[2]School of Veterinary Science, Massey University, Palmerston North 4474, New Zealand
[3]Provita Eurotech Limited, Omagh, County Tyrone, BT79 0EU, Northern Ireland
[4]School of Veterinary Medicine, University of Surrey, Guildford GU2 7AL, United Kingdom
[5]Cattle Lameness Academy, Synergy Farm Health Ltd., Evershot, Dorset, DT2 0LD, United Kingdom
[6]University Farm Animal Practice, 3481 LZ, Harmelen, the Netherlands
[7]School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Leicestershire, LE12 5RD, United Kingdom

ABSTRACT

Digital dermatitis (DD) is the leading infectious cause of lameness in dairy cattle, and it affects their welfare and productivity worldwide. At the herd level, DD is often assessed while cows are standing in a milking parlor, and lesions are most commonly evaluated using the M-score. The objective of this study was to examine the interobserver agreement for M-scores of the feet of standing cattle, based on digital color photographs of dairy cattle hind feet. A total of 88 photographs and written descriptors of the M-score were sent to 11 scorers working at 10 different institutions in 5 countries. The scorers received no formal training immediately before scoring the photographs; however, all regularly used the M-score to score DD. The answers for 36 photographs were excluded from the analysis because the photograph either had more than 1 M-stage as mode or not all scorers assigned an M-score to it. The M-scores of the 11 scorers from 52 photographs were available for analysis. Interobserver agreement was tested using Gwet's agreement coefficient (AC1) and the mode was assumed correct. Overall, moderate agreement emerged for the M-score (AC1 = 0.48). For the individual M-stages, almost perfect agreement existed for M0 (AC1 = 0.99), M1 (AC1 = 0.92), and M3 (AC1 = 0.82), and substantial agreement for M2 (AC1 = 0.61), M4 (AC1 = 0.65), and M4.1 (AC1 = 0.71). This outcome indicates the degree of individual variation in M-scoring in this context by unstandardized, experienced European observers, particularly for the M2, M4, and M4.1 stages. Standardized training is likely to improve the consistency of M-scoring and thus the generalizability of future DD research results on this important endemic disease.

Key words: dairy cow, digital dermatitis, lameness, M-score, interobserver agreement

INTRODUCTION

Bovine digital dermatitis (**DD**) is an endemic infectious disease among farmed cattle. The characteristic active lesion of DD is a painful, large, red to gray ulceration of the skin between the heel bulbs, with the hind feet most often affected. The chronic stage of DD is a dyskeratotic or irregular proliferative hyperkeratotic dermatitis. Despite treatment and control measures, chronic stages often recrudesce into active stages, contributing to further infectious spread of DD and resulting in lameness that compromises animal welfare and productivity (Willshire and Bell, 2009; Bruijnis et al., 2010, 2012).

Several classification systems have been proposed to recognize and grade the visual characteristics of DD lesions. Briefly, Döpfer et al. (1997) classified DD lesions according to morphological observations (**M-score**, which was later adapted by Berry et al., 2012); Laven and Proven (2000) classified DD lesions according to lesion color among other clinical signs; Manske et al. (2002) classified DD lesions according to severity and stage of development; Vink et al. (2009) classified DD lesion according to size, clinical presentation, and location; and Krull et al. (2014) classified DD lesions according to morphological appearance (Iowa-score). Following classification using the M-score, DD lesions were grouped according to disease status as early, infectious, or healing by Döpfer et al. (2012) and as active or inactive by Zinicola et al. (2015) and Biemans et al. (2018).

Received September 4, 2018.
Accepted February 12, 2019.
*Corresponding author: a.vanhoudt@uu.nl

Recognition and grading of DD lesions serves 3 purposes: (1) to study the pathophysiology of DD (Rasmussen et al., 2012; Zinicola et al., 2015; Nielsen et al., 2016), (2) to identify animals that need treatment (Schultz and Capion, 2013; Dotinga et al., 2017), and (3) to study the infection dynamics of DD at a population level (Döpfer et al., 2012; Tremblay et al., 2016; Biemans et al., 2018). Currently, the M-score remains the most widely used, researched, and cited method. Although M-scoring cattle in the trimming chute is considered best practice, regular and repeated screening of herds for DD commonly occurs during a pen walk or milking. Several studies have looked at the diagnostic test characteristics of scoring DD lesions in the milking parlor using various DD lesion classification systems and observations in the trimming chute as the gold standard. For DD lesion classification systems other than the M-score, sensitivity (**Se**) ranges from 65% to 72% and specificity (**Sp**) from 84% to 99% (Rodriguez-Lainz et al., 1998; Thomsen et al., 2008). Yang et al. (2017a) estimated a Se of around 63% and Sp of nearly 100% for visual inspection of the rear feet for presence or absence of lesions of DD during milking. M-scoring in the milking parlor, whether assisted by a telescopic mirror or not, appears to be both sensitive (90–100%) and specific (80–99%) in identifying cattle with DD (Relun et al., 2011; Stokes et al., 2012; Solano et al., 2017), although some misclassification has been reported when compared with M-scoring in the trimming chute, especially for M3 (Relun et al., 2011) and M4.1 (Solano et al., 2017). More recently, Cramer et al. (2018) reported a Se of around 58% and Sp of around 95% after dichotomizing the M-score.

Although the M-score is used by researchers, foot trimmers, farmers, and veterinarians, the methods by which scorers are trained are rarely mentioned in the published literature. In some publications, "an experienced or trained scorer" produces M-scores (Logue et al., 2012; Higginson Cutler et al., 2013; Kulow et al., 2017), whereas elsewhere scorers undergo a detailed training program consisting of recognizing M-stages from color photographs, sometimes followed by scoring live animals (Alsaaod et al., 2014; Solano et al., 2017; Yang et al., 2017b). In the absence of standardized training programs, the reliability and repeatability of DD scoring depends heavily on accurate and consistent interpretation of detailed lesion descriptors written in English. Yet to date, as far as we are aware, the interobserver agreement on M-scoring among scorers working in different institutions has not been studied.

The aim of this study was to assess interobserver agreement of the M-score based on photographs of standing animals. Using several agreement analyses, we calculated the interobserver agreement of the M-score (Döpfer et al., 1997; Berry et al., 2012) among unstandardized, experienced scorers working in different institutions, using single digital color photographs of the hind feet of standing dairy cattle.

MATERIALS AND METHODS

Scorers and Photographs

A convenience sample of 88 digital color photographs of the hind feet (plantar view) of standing dairy cattle was compiled from the personal libraries of 4 scorers (all from the United Kingdom), who respectively contributed 60, 17, 8, and 3 photographs. The photographers were asked to provide photographs from their libraries with high image quality (in focus and taken in a well-lit environment) and absence of lesions other than DD and to include photographs of feet without DD. All but one photograph were marked with ownership. Four photographs were assisted with a telescopic mirror. Nine photographs were annotated with "raised" and 1 with "raised/thickened" features important for scoring that could be seen in real life, but might not be apparent on a photograph. The light source in the photographs varied between natural and artificial light sources including a headlamp. Photographs were taken at varying angles to and distances from the hind feet (estimated range 10–50 cm). The photographed feet were of varying cleanliness. A survey containing the photographs was created in Google forms (Google LLC, Mountain View, CA). The resolution of the original photographs ranged from 1,600 × 1,200 to 3,264 × 1,836 pixels. For compatibility with the Google forms survey, the photographs had to be compressed to resolutions ranging from 269 × 293 to 740 × 991 pixels. An email with the modified M-score descriptors (Table 1; Döpfer et al., 1997; Berry et al., 2012) and a URL (http://bit.ly/M-score_survey) to the survey was sent to 11 scorers, all of whom had scored DD regularly in the past using the M-score. The scorers were asked to complete the survey as they would normally M-score cattle when out on farms. The survey needed to be completed before a certain date, but the time spent on it was not otherwise restricted. The 11 scorers were a convenience sample, without sample size calculation, from within the personal network of the principal investigator. The principal investigator selected the scorers based on them having at least met the proficiency level of the 5-stage model of adult skill acquisition of clinical skills (Dreyfus, 2004). The scorers received no formal training or standardization immediately before the exercise. Scorers could also choose "Don't know" or write a comment for each photograph. Scorers provided the M-scores without interobserver consultation. Upon

Table 1. M-score: M-stage and descriptors, as provided to the scorers

M-stage	Descriptor[1]
M0 or M5[2]	No sign of preexisting lesion. Normal skin.
M1	Small (<2 cm across) focal active state. Circumscribed lesion. Surface is moist, ragged, mottled red–gray with scattered small (~1 mm diameter) red foci.
M2	Larger (>2 cm across) ulcerative active stage. Extensively mottled red–gray. Can be painful upon manipulation.
M3	Healing stage. Typically seen within a few days after antibiotic treatment. The ulcerated surface is now transformed to a dry brown, firm rubbery scab. No pain on manipulation.
M4	Chronic stage. Surface is raised by tan, brown, black, rubbery, irregular, proliferative hyperkeratotic growths that vary from papilliform to mass-like projections.
M4.1	Chronic stage with small active painful M1 focus.

[1]As described by Döpfer et al. (1997) and adapted by Berry et al. (2012).

[2]The M0 stage is more commonly used than the M5 stage described by Berry et al. (2012).

completion of the survey, scorers gave permission to use their data for this research. All 11 scorers answered the survey.

Statistical Analysis

Data were collected by means of Google forms and collated into a spreadsheet (MS Excel, Microsoft Corporation, Redmond, WA). Although M-scores were reported for all photographs, statistical analysis excluded photographs with more than 1 M-stage as mode and photographs not M-scored by all scorers. For each photograph, the mode was assumed the correct M-score. First, the overall mean percentage raw agreements with the mode (**PA_o**; number of exact agreements/total number of observations × 100,) with 95% confidence interval (**CI**) and mean PA_o with 95% CI for each M-stage were calculated. Because the PA_o did not consider the interobserver agreement to be due to chance, we calculated overall Fleiss's kappa (**κ**) with 95% CI (Fleiss, 1971), as well as κ with 95% CI for each M-stage individually. By comparing the PA_o and κ for the individual M-stages, we found a paradox: some M-stages had a high PA_o with a low κ. Therefore, a baseline-category logit model using M-stage as the outcome variable and scorer as a predictor was fitted and the predicted probabilities of reporting each M-stage (category) by each scorer were calculated as follows. Let Y be a nominal outcome with J categories ($J = 1, 2, 3, \ldots, j$) with the probability $\pi_j(X) = P(Y = j|X)$ at a fixed X (the predictor); therefore, $\Sigma_j\pi_j(X) = 1$. Each category J for the outcome Y had probabilities $\{\pi_1(X), \pi_2(X), \ldots, \pi_j(X)\}$. The model relating the probability of category j to that of a baseline category (for example, $J = 1$) could then be formulated as

$$\ln\left(\frac{\pi_j(X)}{\pi_1(X)}\right) = \beta_0^{(j)} + \beta_1^{(j)}X,$$

where β_0 is the intercept and β_1 measures the effect of scorers for each of the J categories. The predicted probability for any scorer reporting any category was

$$\pi_j(X) = \frac{e^{\beta_0^{(j)}+\beta_1^{(j)}X}}{1+\sum_{j=1}^{J} e^{\beta_0^{(j)}+\beta_1^{(j)}X}}.$$

The variances (σ^2) of the predicted probabilities for each M-stage were used as indicators to describe the variability across scorers for each M-stage. This approach revealed that the high PA_o together with a low κ for some M-stages was due to unequal prevalence (based on the mode) of the M-stages in our data set. Finally, for more robust and relevant measurement of interobserver agreement, Gwet's agreement coefficient (**AC1**; Gwet, 2008) was used as it is less sensitive to either marginal homogeneity or trait prevalence. Gwet's AC1 with 95% CI was calculated for overall agreement and each M-stage separately. We recalculated Gwet's AC1 with 95% CI for overall agreement after condensing several M-stages into different groups (Table 2). The analysis of the baseline-category logit model was done using Stata 13.1 (StataCorp LLC, College Station, TX), and all other statistical analyses were done using R (R Core Team, 2014).

For all measures of agreement, the guidance provided by Landis and Koch (1977) for the interpretation of κ was used: <0.00, poor; 0.00 to 0.20, slight; 0.21 to 0.40, fair; 0.41 to 0.60, moderate; 0.61 to 0.80, substantial; and 0.81 to 1.00, almost perfect.

RESULTS

Scorers and Photographs

The 11 scorers were geographically distributed over England (7), the Netherlands (1), Northern Ireland (1),

Table 2. Overview of the groups of M-stages used for Gwet's agreement coefficient (AC1) calculation

	M-stage[1]					
Grouping criterion	M0[2]	M1	M2	M3	M4	M4.1
Lesion color (Laven and Proven, 2000)	No lesion	Red	Red	Black	Gray	Gray
Infectious disease modeling by Döpfer et al. (2012)	No lesion	Early	Infectious	Healing	Infectious	Infectious
Infectious disease modeling by Biemans et al. (2018)	No lesion	Active	Active	Inactive	Inactive	Active
Absence or presence of digital dermatitis	No lesion	Lesion	Lesion	Lesion	Lesion	Lesion

[1]As described by Döpfer et al. (1997) and adapted by Berry et al. (2012).
[2]The M0 stage is more commonly used than the M5 stage described by Berry et al. (2012).

the Republic of Ireland (1), and Spain (1). Six scorers were employed by 5 different universities, 2 by different agricultural companies, and 3 were self-employed veterinary consultants. Ten scorers held a degree in veterinary medicine and 1 scorer in agri-food and business studies. Most of the scorers (9) also held at least 1 postgraduate degree. At the moment of answering the survey, 2 scorers were senior researchers in the field of bovine lameness; 2 scorers had recently obtained a PhD in a relevant field; 2 scorers were PhD candidates in a relevant field; 2 scorers were residents of the European College of Bovine Health Management; 2 scorers were in a commercial role, with one having obtained a PhD on digital dermatitis; and 1 scorer was a farm animal veterinary consultant. Between the scorers, experience in using the M-score varied, with 6 scorers having 1 to 5 years of experience, 4 scorers having 6 to 10 years of experience, and 1 scorer having 16 to 20 years of experience.

All but 1 scorer assessed all the photographs. One scorer could not assess 3 photographs due to an error in opening them and 1 photograph received a blank response from this scorer. Another scorer gave the general comment "The diagnosis of M1 and M3 is limited from pictures as M1 is difficult to spot and M3s by definition occur as a transitory state after treatment." That scorer did not assign any photograph with the M1 or M3 stage. The number of photographs assigned an M-stage by each scorer ranged from 76 to 88, with 4 scorers assigning an M-stage to all 88 photographs. Table 3 summarizes the assigned M-scores and the modes for the 88 photographs. The answers for 6 (7%) photographs were excluded because they had more than 1 M-stage as mode (e.g., photograph 30 was scored as M3 by 5 scorers, M4 by 5 scorers, and M4.1 by 1 scorer) and the answers for 30 (34%) photographs were excluded because they did not receive an M-stage from all scorers (21 photographs were not given an M-stage by 1 scorer, 3 by 2 scorers, 3 by 3 scorers, and 3 by 4 scorers). The M-scores for 52 (59%) photographs were used for analysis. The resolution after compression for the survey was 740 × 555 pixels for the 6 photographs excluded for having more than 1 M-stage as mode, ranged from 269 × 293 to 740 × 991 pixels for the 30 photographs excluded for not receiving an M-stage from all scorers, and ranged from 505 × 367 to 740 × 991 pixels for the 52 photographs used for analysis.

Agreement Analyses

At the level of the scorer, mean PA_o (95% CI) was 72% (64–79%) and mean PA_o at the level of the photo-

Table 3. Descriptive data showing the M-scores[1] assigned to 88 digital color photographs of the hind feet of standing dairy cattle by 11 experienced but unstandardized scorers; the frequencies of "correct" (mode[2]; bold) and other classifications are shown, both for the number of M-scores assigned and the number of photographs that the scores were assigned to

	Actual classification, count of scores given (count of photographs)					
Item	M0[3]	M1	M2	M3	M4	M4.1
M0	**93 (10)**	3 (3)	1 (1)	1 (1)	5 (4)	0 (0)
M1	1 (1)	**18 (4)**	2 (2)	4 (3)	7 (3)	3 (2)
M2	0 (0)	14 (8)	**158 (22)**	13 (6)	11 (6)	42 (19)
M3	2 (1)	18 (9)	8 (6)	**57 (14)**	52 (14)	3 (3)
M4	6 (3)	28 (15)	25 (13)	100 (42)	**277 (45)**	27 (17)
M4.1	0 (0)	1 (1)	20 (7)	7 (4)	11 (5)	**48 (8)**

[1]As described by Döpfer et al. (1997) and adapted by Berry et al. (2012).
[2]The mode (bold type) was taken to be the correct classification. For 11 photographs, there were 2 modes, and for 2 photographs, there were 3 modes.
[3]The M0 stage is more commonly used than the M5 stage described by Berry et al. (2012).

Table 4. Statistical analyses for agreement with the mode with 95% CI for each M-stage[1] and overall for the M-score[1] from 52 digital color photographs of the hind feet of standing dairy cattle assessed by 11 experienced but unstandardized scorers

Variable	Overall	M0[2]	M1	M2	M3	M4	M4.1
N[3]	52	6	1	19	1	19	6
Percent raw agreement (95% CI)	72 (67–76)	97 (94–100)	73[4]	68 (61–75)	55[4]	71 (64–77)	61 (51–71)
Fleiss's κ (95% CI)	0.44*** (0.36–0.53)	0.96*** (0.92–1.00)	0.23* (0.00–0.48)	0.45*** (0.35–0.56)	0.10** (0.02–0.18)	0.51*** (0.40–0.61)	0.23*** (0.12–0.34)
Variance[5]	—	0.000	0.003	0.013	0.005	0.007	0.014
Gwet's agreement coefficient, AC1 (95% CI)	0.48*** (0.41–0.56)	0.99*** (0.98–1.00)	0.92*** (0.88–0.97)	0.61*** (0.46–0.75)	0.82*** (0.74–0.89)	0.65*** (0.51–0.79)	0.71*** (0.61–0.82)

[1]As described by Döpfer et al. (1997) and adapted by Berry et al. (2012).
[2]The M0 stage is more commonly used than the M5 stage described by Berry et al. (2012).
[3]Number of photographs (with this M-stage as the mode).
[4]No 95% CI for the mean percentage raw agreement with the mode for single observations.
[5]Variances of the predicted probabilities of reporting each M-stage by each scorer following baseline-category logit model analysis.
*$P = 0.08$, **$P = 0.01$, ***$P < 0.001$ (within rows).

graph was also 72% (67–76%). We found 100% agreement for only 5 (10%) photographs (4 M0 and 1 M4) and at least 60% agreement for 40 (77%) photographs. For each M-stage and overall for the M-score, the results of the statistical agreement analyses (i.e., PA_o, κ, σ^2, and AC1) are given in Table 4. After grouping the M-stages, the overall AC1 (95% CI) for the M-score was 0.56 (0.49 to 0.64, $P < 0.001$) for lesion color as used by Laven and Proven (2000), 0.74 (0.67 to 0.81, $P < 0.001$) for infectious disease modeling classification as used by Döpfer et al. (2012), 0.78 (0.71 to 0.86, $P < 0.001$) for infectious disease modeling classification as used by Biemans et al. (2018), and 0.99 (0.98 to 1.00, $P < 0.001$) for absence or presence of a DD lesion.

DISCUSSION

This study demonstrates the variation in agreement between users when M-scoring digital color photographs of the hind feet of standing dairy cattle. Overall, mean PA_o was around 70% at the level of the photograph. The PA_o between observers for the individual M-stages was moderate (M3), substantial (M1, M2, M4, and M4.1), or almost perfect (M0). Fleiss's κ analysis highlights that agreement is poorer when adjusted for agreement due to chance (slight for M3, fair for M1 and M4.1, moderate for M2 and M4, and almost perfect for M0). Using Gwet's AC1, which accounts for marginal homogeneity and trait prevalence, we found an improvement in the interobserver agreement for all M-stages when compared with κ agreement (substantial for M2, M4, and M4.1, and almost perfect for M0, M1, and M3). The overall AC1 agreement for the M-score improved in comparison with overall κ agreement (κ = 0.44) but remained only moderate (AC1 = 0.48).

Few studies have looked at interobserver agreement of the M-score (Relun et al., 2011; Biemans et al., 2018; Solano et al., 2017) and of these, only 1 describes the interobserver agreement of the M-score when applied to digital color photographs of hind feet (Solano et al., 2017) (Table 5). In these studies, Cohen's κ is used to measure interobserver agreement (Cohen, 1960). This study is the first using Fleiss's κ and Gwet's AC1 to investigate interobserver agreement of the M-score when applied to digital color photographs of hind feet, thereby accounting for having more than 2 observers, marginal homogeneity, trait prevalence, and agreement due to chance with a more reasonable assumption. It is impossible to know what the interobserver agreement would have been in the other studies had they used Fleiss's κ, Gwet's AC1, or both, which impedes interpreting the results from this study in light of those from previous studies.

Care should be taken in comparing the interobserver agreement of the M-score from digital color photographs of cattle feet with those from studies using live animals. Digital color photographs show the feet in a set 2-dimensional view (versus a changeable 3-dimensional view in real life), which makes estimating the dimensions of the lesion difficult and thereby limits the observer's ability to interpret the presented foot. Also, certain aspects of the M-score descriptors cannot be considered when scoring from photographs (i.e., reaction on manipulation and treatment history). However, these difficulties apply equally to the screening of standing animals during pen walks or in the milking parlor. Further, treatment history is not clearly stated as an essential criterion in the M-score descriptors, and only 1 scorer commented that treatment history was unknown; this scorer consequently did not assign the

M3 stage (or with the M1 stage) to any photograph. This scorer also did not assign an M-stage to 4 photographs. The mode for these 4 photographs was neither M1 nor M3. Because scorers did not explain why they did not to assign an M-score to a photograph, the true reason for not assigning an M-score to any photograph is unknown. Future DD research using photographs of cattle feet should alleviate the limitations of M-scoring photographs as much as possible by using novel image capture techniques to resemble human vision (e.g., stereo-vision capture systems) or including a ruler in the photograph and using photographs taken under standard conditions, that is, using the same camera under the same lighting conditions, and taken by the same photographer at the same distance and angle to the foot.

The advantages of M-scoring from photographs are that the animals do not move and there is no time pressure, unlike when M-scoring live animals during milking. It also allows more effective blinding of observers, thereby accounting for observer drift. Using photographs of cattle feet for DD research offers the opportunity to amass scorers from a large population of researchers for international standardization, with guidance on interpretation from the most experienced and competent scorers or a remotely located expert scorer.

The level of interobserver agreement in our study is lower than that reported by others regardless of whether they scored digital color photographs or live animals. One possible reason for the lower interobserver agreement in this study is the lack of the prestudy training, which was provided in some other studies (Relun et al., 2011; Solano et al., 2017; Biemans et al., 2018). As far as we are aware, this study is the first to assess the M-score interobserver agreement with observers from (10) different institutions. This factor may have contributed to the lower interobserver agreement in this study compared with previous studies using observers working in the same institution (Relun et al., 2011; Solano et al., 2017; Biemans et al., 2018). Future research is needed to confirm this possibility. Because the diversity of the scorers in our study may have contributed to the difference in interobserver agreement, it may also cast doubt over the comparability of international DD research. It is possible that scorer characteristics, such as sex, age, type of qualification, years of experience in applying the M-score, and the method of training in DD scoring,

Table 5. Overview of interobserver agreement statistics with 95% CI for the M-score[1] from this study and those found in the published literature

Study object and study	N scorers	M-score[1]	Experimental units	Interobserver agreement statistic (95% CI)		
				Percent raw agreement[2]	Kappa	Gwet's agreement coefficient[3]
Photographs						
This study	11	6 Stage	52 digital color photographs of hind feet taken from standing dairy cattle (plantar view)	72 (67–76)	Fleiss[4]: 0.41 (0.33–0.49)	0.48 (0.41–0.56)
Solano et al., 2017	3	6 Stage	40 digital color photographs of hind feet (start study)	83 (70–94)	Cohen[5]: 0.77 (0.67–0.86)	—
			40 digital color photographs of hind feet (midway study)	88 (76–98)	Cohen: 0.83 (0.74–0.90)	—
Live animals						
Relun et al., 2011	5	5 Stage	Hind feet from 242 cows in the milking parlor	66 (62–70)	Cohen: 0.51 (0.45–0.56)	—
Solano et al., 2017	3	6 Stage	Hind feet from 110 cows in the milking parlor	82 (73–90)	Cohen: 0.74 (0.69–0.78)	—
Biemans et al., 2018	2	6 Stage	204 hind feet in the milking parlor (start study)	—	Cohen: 0.75 (0.66–0.84)	—
			164 hind feet in the milking parlor (during study)	—	Cohen: 0.85 (0.78–0.93)	—
			52 hind feet in the milking parlor (during study)	—	Cohen: 0.76 (0.61–0.90)	—

[1]As described by Döpfer et al. (1997) and adapted by Berry et al. (2012); the M0 stage is more commonly used than the M5 stage described by Berry et al. (2012); 5-stage classification (M0, M1, M2, M3, and M4) or 6-stage classification (M0, M1, M2, M3, M4, and M4.1).

[2]Percent raw agreement is the number of exact agreements divided by the total number of observations multiplied by 100.

[3]Gwet's agreement coefficient is suitable for multiclass, multi-observer interobserver agreement analysis. It corrects for agreement due to chance with a more reasonable assumption and thus is less sensitive to either marginal homogeneity or trait prevalence (Gwet, 2008).

[4]Fleiss's kappa is suitable for multiclass, multi-observer interobserver agreement analysis and corrects for agreement due to chance (Fleiss, 1971).

[5]Cohen's kappa is suitable for multiclass interobserver agreement analysis of 2 observers and corrects for agreement due to chance (Cohen, 1960).

could influence interobserver agreement. Unfortunately, these influences could not be investigated in our study.

We did find that grouping the M-stages resulted in higher AC1 agreements. Grouping certain M-stages, as both Relun et al. (2011) and Solano et al. (2017) found, yields higher interobserver agreement. In this study, dichotomizing the M-score as absent or present resulted in the highest overall AC1 interobserver agreement (0.99). This is also reflected in the almost perfect agreement between the scorers for the photographs with M0 as the mode in this study, regardless of the type of statistical agreement analysis. We interpret this finding as implying that all scorers are generally well able to identify cattle with and without DD on digital color photographs of the hind feet of standing dairy cattle. Further research is needed to identify which M-stages should be grouped for each type of use (pathophysiology, treatment, or infection dynamics of DD) and scorer (researcher, foot trimmer, farmer, or veterinarian) to enable highest interobserver agreement, while maintaining sufficient diagnostic test characteristics such as sensitivity and specificity.

In our data set, 30 photographs were not assigned an M-stage by every scorer, meaning that at least 1 scorer was unsure which M-stage the photograph represented. This was likely to be a consequence of lesion descriptor interpretation, photograph limitations (versus real life), lesion complexity, the standing position of the leg (versus inspecting raised feet in the trimming chute), or a combination of these factors. Unfortunately, during data collection scorers were not asked to give their reason for not assigning an M-stage to a photograph. Excluding these 30 photographs and the 6 photographs with more than 1 M-stage as the mode likely caused a bias toward the best quality photographs because all scorers were presumably confident about their M-scores for the remaining 52 photographs that were used for agreement analysis.

Achieving high interobserver agreement for the recognition and classification of DD lesions is crucial for international generalizability and applicability of the results from DD research. The development of an internationally available DD training program would likely help in achieving high interobserver agreement for the recognition and classification of DD lesions, although this outcome should be confirmed in future research. Any future DD training program should take into account the intended use of the classification system (pathophysiology, treatment, or infection dynamics of DD) and user type (researcher, foot trimmer, farmer, or veterinarian). In addition, the application of automated DD lesion recognition and classification using novel image capturing techniques and artificial intelligence should be researched and developed. This approach would enable early cow-side diagnosis of cattle eligible for treatment and disease status monitoring, both on farms with automated milking systems and on farms with conventional milking systems.

CONCLUSIONS

The aim of this study was to investigate the interobserver agreement of the M-score applied to digital color photographs of the hind feet of standing dairy cattle when scored by observers working in different institutions. We studied the external validity of the M-score, which reflects the generalizability of the results from DD research using the M-score. The results from this study indicate that the external validity of the M-score is almost perfect when dichotomized as the absence or presence of a DD lesion but lower for the M2, M4, and M4.1 stages, the 3 stages that are assigned important roles in the clinical aspect or epidemiology of DD. Achieving high interobserver agreement for all the M-stages between scorers globally would greatly benefit the investigation of DD because it will contribute to the comparability of future DD research results. We propose that standardized training of scorers would likely improve the consistency between scorers, and this possibility should be the focus of future research.

ACKNOWLEDGMENTS

The authors acknowledge D. Döpfer (School of Veterinary Medicine, University of Wisconsin-Madison, USA) for her feedback on the study and A. Gomez (Zinpro, Spain) for scoring the photographs and his feedback on the study. We thank the R. Blowey, S. Pedersen (Farm Dynamics Ltd., United Kingdom), J. Somers (School of Veterinary Medicine, University College Dublin, the Republic of Ireland), and J. Stokes (School of Veterinary Sciences, University of Bristol, United Kingdom) for their time and efforts scoring the photographs, and Linda McPhee (Linda McPhee Consulting, United Kingdom) for language editing of the manuscript. R. Blowey and S. Pedersen also provided photographs for the survey. Preliminary results of this study were presented by NJB at the 19th International Symposium and 11th Conference on Lameness in Ruminants (September 2017, Munich, Germany) and by AV at the 20th International Symposium and 12th Conference on Lameness in Ruminants (March 2019, Tokyo, Japan). At the time of the study, ADM was employed by The Royal Veterinary College (United Kingdom) and RFN and JNH by the School of Veterinary Medicine and Science (University of Nottingham, United Kingdom).

AV scored the photographs, prepared the dataset for analysis, and wrote the manuscript, all other authors contributed to the manuscript; DAY performed the statistical analysis; TA provided photographs, made the survey, scored the photographs, and compiled the dataset; ADM and RFN scored the photographs; NJB designed the study, provided photographs, and scored the photographs. TA was employed by Provita Eurotech Limited (Northern Ireland) at the time of the study and is currently employed by Lely Center Eglish (Northern Ireland). RFN is affiliated with the Cattle Lameness Academy of Synergy Farm Health Ltd. (United Kingdom). TvW is affiliated with the University Farm Animal Practice (the Netherlands). NJB is founder and owner of Bos International Ltd. (United Kingdom). All authors declare that they have no conflict of interest related to the study discussed in this manuscript.

REFERENCES

Alsaaod, M., C. Syring, J. Dietrich, M. G. Doherr, T. Gujan, and A. Steiner. 2014. A field trial of infrared thermography as a non-invasive diagnostic tool for early detection of digital dermatitis in dairy cows. Vet. J. 199:281–285. https://doi.org/10.1016/j.tvjl.2013.11.028.

Berry, S. L., D. H. Read, T. R. Famula, A. Mongini, and D. Döpfer. 2012. Long-term observations on the dynamics of bovine digital dermatitis lesions on a California dairy after topical treatment with lincomycin HCl. Vet. J. 193:654–658. https://doi.org/10.1016/j.tvjl.2012.06.048.

Biemans, F., P. Bijma, N. M. Boots, and M. C. M. de Jong. 2018. Digital dermatitis in dairy cattle: The contribution of different disease classes to transmission. Epidemics 23:76–84. https://doi.org/10.1016/j.epidem.2017.12.007.

Bruijnis, M. R. N., B. Beerda, H. Hogeveen, and E. N. Stassen. 2012. Assessing the welfare impact of foot disorders in dairy cattle by a modeling approach. Animal 6:962–970. https://doi.org/10.1017/S1751731111002606.

Bruijnis, M. R. N., H. Hogeveen, and E. N. Stassen. 2010. Assessing economic consequences of foot disorders in dairy cattle using a dynamic stochastic simulation model. J. Dairy Sci. 93:2419–2432. https://doi.org/10.3168/jds.2009-2721.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37–46. https://doi.org/10.1177/001316446002000104.

Cramer, G., T. Winders, L. Solano, and D. H. Kleinschmit. 2018. Evaluation of agreement among digital dermatitis scoring methods in the milking parlor, pen, and hoof trimming chute. J. Dairy Sci. 101:2406–2414. https://doi.org/10.3168/jds.2017-13755.

Döpfer, D., M. Holzhauer, and M. van Boven. 2012. The dynamics of digital dermatitis in populations of dairy cattle: model-based estimates of transition rates and implications for control. Vet. J. 193:648–653. https://doi.org/10.1016/j.tvjl.2012.06.047.

Döpfer, D., A. Koopmans, F. A. Meijer, I. Szakáll, Y. H. Schukken, W. Klee, R. B. Bosma, J. L. Cornelisse, A. J. van Asten, and A. A. ter Huurne. 1997. Histological and bacteriological evaluation of digital dermatitis in cattle, with special reference to spirochaetes and *Campylobacter faecalis*. Vet. Rec. 140:620–623. https://doi.org/10.1136/VR.140.24.620.

Dotinga, A., R. Jorritsma, and M. Nielen. 2017. A randomised non-inferiority trial on the effect of an antibiotic or non-antibiotic topical treatment protocol for digital dermatitis in dairy cattle. Vet. Evid. 2. https://doi.org/10.18849/ve.v2i4.111.

Dreyfus, S. E. 2004. The five-stage model of adult skill acquisition. Bull. Sci. Technol. Soc. 24:177–181. https://doi.org/10.1177/0270467604264992.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76:378–382. https://doi.org/10.1037/h0031619.

Gwet, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61:29–48. https://doi.org/10.1348/000711006X126600.

Higginson Cutler, J. H., G. Cramer, J. J. Walter, S. T. Millman, and D. F. Kelton. 2013. Randomized clinical trial of tetracycline hydrochloride bandage and paste treatments for resolution of lesions and pain associated with digital dermatitis in dairy cattle. J. Dairy Sci. 96:7550–7557. https://doi.org/10.3168/jds.2012-6384.

Krull, A. C., J. K. Shearer, P. J. Gorden, V. L. Cooper, G. J. Phillips, and P. J. Plummer. 2014. Deep sequencing analysis reveals temporal microbiota changes associated with development of bovine digital dermatitis. Infect. Immun. 82:3359–3373. https://doi.org/10.1128/IAI.02077-14.

Kulow, M., P. Merkatoris, K. S. Anklam, J. Rieman, C. Larson, M. Branine, and D. Döpfer. 2017. Evaluation of the prevalence of digital dermatitis and the effects on performance in beef feedlot cattle under organic trace mineral supplementation. J. Anim. Sci. 95:3435–3444. https://doi.org/10.2527/jas.2017.1512.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310.

Laven, R. A., and M. J. Proven. 2000. Use of an antibiotic footbath in the treatment of bovine digital dermatitis. Vet. Rec. 147:503–506. https://doi.org/10.1136/VR.147.18.503.

Logue, D. N., T. Gibert, T. Parkin, S. Thomson, and D. J. Taylor. 2012. A field evaluation of a footbathing solution for the control of digital dermatitis in cattle. Vet. J. 193:664–668. https://doi.org/10.1016/j.tvjl.2012.06.050.

Manske, T., J. Hultgren, and C. Bergsten. 2002. Topical treatment of digital dermatitis associated with severe heel-horn erosion in a Swedish dairy herd. Prev. Vet. Med. 53:215–231. https://doi.org/10.1016/S0167-5877(01)00268-9.

Nielsen, M. W., M. L. Strube, A. Isbrand, W. D. H. M. Al-Medrasi, M. Boye, T. K. Jensen, and K. Klitgaard. 2016. Potential bacterial core species associated with digital dermatitis in cattle herds identified by molecular profiling of interdigital skin samples. Vet. Microbiol. 186:139–149. https://doi.org/10.1016/j.vetmic.2016.03.003.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rasmussen, M., N. Capion, K. Klitgaard, T. Rogdo, T. Fjeldaas, M. Boye, and T. K. Jensen. 2012. Bovine digital dermatitis: Possible pathogenic consortium consisting of *Dichelobacter nodosus* and multiple *Treponema* species. Vet. Microbiol. 160:151–161. https://doi.org/10.1016/j.vetmic.2012.05.018.

Relun, A., R. Guatteo, P. Roussel, and N. Bareille. 2011. A simple method to score digital dermatitis in dairy cows in the milking parlor. J. Dairy Sci. 94:5424–5434. https://doi.org/10.3168/jds.2010-4054.

Rodriguez-Lainz, A., P. Melendez-Retamal, D. W. Hird, and D. H. Read. 1998. Papillomatous digital dermatitis in Chilean dairies and evaluation of a screening method. Prev. Vet. Med. 37:197–207. https://doi.org/10.1016/S0167-5877(98)00091-9.

Schultz, N., and N. Capion. 2013. Efficacy of salicylic acid in the treatment of digital dermatitis in dairy cattle. Vet. J. 198:518–523. https://doi.org/10.1016/j.tvjl.2013.09.002.

Solano, L., H. W. Barkema, C. Jacobs, and K. Orsel. 2017. Validation of the M-stage scoring system for digital dermatitis on dairy cows in the milking parlor. J. Dairy Sci. 100:1592–1603. https://doi.org/10.3168/jds.2016-11365.

Stokes, J. E., K. A. Leach, D. C. J. Main, and H. R. Whay. 2012. The reliability of detecting digital dermatitis in the milking parlour. Vet. J. 193:679–684. https://doi.org/10.1016/j.tvjl.2012.06.053.

Thomsen, P. T., I. C. Klaas, and K. Bach. 2008. Short communication: Scoring of digital dermatitis during milking as an alternative

to scoring in a hoof trimming chute. J. Dairy Sci. 91:4679–4682. https://doi.org/10.3168/jds.2008-1342.

Tremblay, M., T. Bennett, and D. Döpfer. 2016. The DD Check App for prevention and control of digital dermatitis in dairy herds. Prev. Vet. Med. 132:1–13. https://doi.org/10.1016/j.prevetmed.2016.07.016.

Vink, W. D., G. Jones, W. O. Johnson, J. Brown, I. Demirkan, S. D. Carter, and N. P. French. 2009. Diagnostic assessment without cut-offs: Application of serology for the modelling of bovine digital dermatitis infection. Prev. Vet. Med. 92:235–248. https://doi.org/10.1016/j.prevetmed.2009.08.018.

Willshire, J. A., and N. J. Bell. 2009. An economic review of cattle lameness. Cattle Pract. 17:136–141.

Yang, D. A., C. Heuer, R. Laven, W. D. Vink, and R. N. Chesterton. 2017a. Estimating the true prevalence of bovine digital dermatitis in Taranaki, New Zealand using a Bayesian latent class model. Prev. Vet. Med. 147:158–162. https://doi.org/10.1016/j.prevetmed.2017.09.008.

Yang, D. A., C. Heuer, R. Laven, W. Vink, and R. Chesterton. 2017b. Farm and cow-level prevalence of bovine digital dermatitis on dairy farms in Taranaki, New Zealand. N. Z. Vet. J. 65:252–256. https://doi.org/10.1080/00480169.2017.1344587.

Zinicola, M., F. Lima, S. Lima, V. Machado, M. Gomez, D. Döpfer, C. Guard, and R. Bicalho. 2015. Altered microbiomes in bovine digital dermatitis lesions, and the gut as a pathogen reservoir. PLoS One 10:e0120504. https://doi.org/10.1371/journal.pone.0120504.