Validation of an automatic scoring system for the assessment of hock burn in broiler

Helen Louton ,*,[1] Andre Piller,[†] Shana Bergmann,[†] Michael Erhard,[†] Paul Schmidt ,[‡] Nicole Kemper ,[§] Jan Schulte-Landwehr,[#] and Angela Schwarzer[†]

[*]Animal Health and Animal Welfare, Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock D-18059, Germany; [†]Chair of Animal Welfare, Animal Behaviour, Animal Hygiene and Animal Husbandry, Department of Veterinary Sciences, Faculty of Veterinary Medicine, LMU Munich, Munich D-80539, Germany; [‡]Paul Schmidt, Statistical Consulting for Science and Research, Berlin D-13086, Germany; [§]Institute for Animal Hygiene, Animal Welfare and Farm Animal Behaviour, University of Veterinary Medicine Hannover, Foundation, Hannover D-30173, Germany; and [#]CLK GmbH, Altenberge D-48341, Germany

ABSTRACT This study aimed to develop and validate a camera vision score that could detect macroscopic alterations of the hock, to identify errors and to assess possible factors that could influence the assessment. Two hundred hocks in the first (calibration) phase and 500 hocks in the second (validation) phase were collected at slaughter, visually assessed, placed back into the evisceration line and assessed by a camera system with 2 software systems. The size of the alteration in percent (%) measured by the camera system was evaluated ("camera score", **CS**). Additionally, temperature, humidity, and light intensities were measured. In the calibration phase, threshold values of camera scores for respective macro scores were defined and performance measures evaluated. In the validation phase, the generated threshold values were validated, occurring errors, as well as possible impacts of climatic factors analyzed. The results showed that the generated thresholds predict the camera score values at which the respective macro score has the highest probability of appearance. Small hock burn lesions ≤0.5 cm have the highest probability at a camera score of ≥0.2 (original CS) or ≥0.1 (updated CS), and lesions >0.5 cm have the highest probability at a camera score of ≥0.7 (original CS) or ≥1.1 (updated CS). Large lesions (>0.5 cm) are more reliably identified by the system than small lesions. The risks of errors in assessing reference areas and lesions showed a correct identification of lesions to be the most probable result even if the reference area is not correctly identified. The probability of a correct identification of lesions by the camera system was slightly higher (not significant) with the updated software (risk = 0.66 [0.62−0.70]) than with the original software (risk = 0.63 [0.58−0.67]). Automatic assessment systems at slaughter could be adjusted to the presented threshold values to classify hock burn lesions. Software adaptations can improve the performance measures of diagnosis and reduce the probability of errors.

Key words: automatic assessment, broiler, precision livestock farming, welfare indicator, hock burn

2022 Poultry Science 101:102025
https://doi.org/10.1016/j.psj.2022.102025

INTRODUCTION

Currently, there is increasing attention to evaluate welfare of poultry via the assessment of animal-based indicators rather than merely evaluating the housing or management conditions (Butterworth, 2013). Johnsen et al. (2001) describe nine methods to identify animal welfare via herd-level assessment. The parameters are classified into at least 2 types: animal-based and environment-based indicators. Animal-based indicators, such as the health of animals, are directly measurable parameters and assess the animals' reaction to their specific environment. However, the assessment of animal-based indicators is rather time consuming (Johnsen et al., 2001). In the scope of the European Welfare Quality Project, standardized methods to evaluate animal welfare were developed (Welfare Quality® Network, 2018). In the Welfare Quality® assessment protocol for poultry (2009), 4 welfare principles were split into 12 welfare criteria. For poultry for example, animal-based indicators were defined in such a way that broiler welfare could be assessed on farm and at slaughter. Among

Received April 8, 2022.
Accepted June 18, 2022.

[1]Corresponding author: helen.louton@uni-rostock.de

others, foot pad dermatitis and hock burn were identified as suitable indicators for the assessment of the welfare on-farm and at slaughter. In one third of the member states of the European Union, representing 32% of the broiler production, the prevalence and severity of foot pad dermatitis are routinely assessed at slaughter of the animals. Further member states are planning to implement this assessment (European Union, 2016). Slaughterhouses seem to be particularly suited for the assessment of these animal welfare indicators because a large number of animals can be assessed with a comparably low effort (Louton et al., 2018a; Jung et al., 2021). In Germany, the official veterinarian at the slaughterhouse is obliged to assess and to report welfare indicators indicative of unsuited housing conditions (Maisack, 2016).

Hock burn, a contact dermatitis of the skin of the hock of broilers (Greene et al., 1985; Bessei, 2006), is an animal-based indicator and suitable to be assessed at slaughter because it represents the health and welfare of the birds (Hepworth et al., 2011; Saraiva et al., 2016). Hock burn is a common welfare issue in broilers and prevalence ranges from 35% (fattening d 35; Bergmann et al., 2016) to 88% (wk 6; Kjaer et al., 2006) of affected birds, showing large inter flock variations. Haslam et al. (2007) found that on average 1.3% of the birds had moderate or severe hock burn with a range of 0 to 33%, in contrast in each flock 11% of the birds had moderate or severe FPD with a range of 0 to 72%. Louton et al. (2018b) observed higher hock burn rates than FPD rates in conventionally housed broilers.

Even though hock burn is a suitable indicator to assess the welfare of the birds, no standardized governmental scoring system to assess these lesions at slaughter is available in Europe. Scoring systems reported by scientists include a scoring system by McKeegan (2010), the Welfare Quality assessment protocol for poultry (Welfare Quality®, 2009) and the assessment scheme recently recommended by our research group, who performed a histological validation of a macroscopic hock burn scoring system (Louton et al., 2020a).

The assessment of animal-based indicators via a camera system is one possibility of evaluating animal welfare. Ben Sassi et al. (2016) summarized that precision livestock farming technologies provide valuable information for farmers, allow an improvement of welfare and environmental impact, and can be used as efficient and early detection tools. Several authors assessed animal health or welfare on farm by automatic assessment systems (Aydin, 2017; Dawkins et al., 2017; Zhuang et al., 2018). Advantages of a camera vision assessment are that all animals of a flock can be assessed, the needed workforce can be reduced, and the reliability of assessments can potentially be increased by standardization (Jung et al., 2021). However, Vanderhasselt et al. (2013) showed that the assessment of foot pad dermatitis of broilers via a camera system (using computer vision) was to some extent not correct, and the scores generated by the automatic assessment correlated only weakly with the scores evaluated by experts. Yet, according to these authors, the automatic assessment of animal welfare indicators seemed promising. Once the definitely incorrect assessments were identified and discarded, the automatic system seemed not to deviate strongly from the assessment by auditing experts (Vanderhasselt et al., 2013). The automatic assessment of foot pad dermatitis was validated by Louton et al. (2022) and showed that, in particular no lesions and larger foot pad dermatitis lesions (>0.5 cm) were identified with a high sensitivity by a camera vision based system. However, especially small lesions (≤0.5 cm) were not identified with a sufficient sensitivity (Louton et al., 2022). The automatic assessment of hock burn in broilers is currently performed at several slaughter plants in Germany. In a small-scaled study, one of 3 interviewed slaughter plant operators mentioned to assess hock burn as a welfare indicator, and assessed this indicator automatically via a camera system (Louton et al., 2018a). However, to our knowledge, literature considering threshold values and the validity of these kinds of automatic assessment systems is lacking, especially for hock burn lesions.

Thus, automatic assessment systems are promising for evaluating welfare conditions in livestock and can be used to reflect the health of animals. However, threshold values need to be chosen carefully and should be validated to provide reliable results. Software adaptations could be used to improve the performance measures of diagnosis and to reduce the probability of errors during automatic assessment. The general goal of this study is to find improved options for automatic assessment systems that could support camera vision assessment as a tool to evaluate animal welfare indicators (hock burn as indicator). Thereby the following objectives were defined: to develop a camera vision score that could detect macroscopic alterations of the hock: 1) to validate this camera vision scoring system for the classification of hock burn in broilers, 2) and to identify errors, 3) and assess possible influencing external factors, 4) when reference surface areas and lesions of hock burns were considered.

ANIMALS, MATERIALS, AND METHODS

Animals and Material

For the project, hocks from Ross 308 broilers of an age of 36 to 42 fattening days were collected at a poultry slaughterhouse in Bavaria, Germany.

Interobserver Reliability Test

To assess the interobserver agreement, the interobserver reliability test was performed using the prevalence-adjusted and bias-adjusted kappa (**PABAK**) calculation by Byrt et al. (1993) for the assessment of 250 hocks of Ross 308 broilers, evaluated by 5 observers. The procedure and results of the interobserver reliability test were published by Louton et al. (2020a), who conducted a study in which the used macroscopic assessment scheme was histologically validated. Two of the 5 observers performed all assessments in the first and second phase of the presented study. Training was done prior to the assessment of the first phase by schooling untrained observers. Consultation and interobserver agreement were done regularly, since assessments were commonly done together by the 2 observers (of which one was experienced).

Visual Assessment

Phase 1 (Calibration) Initially, 200 hocks (40 hocks of each macroscopic score) from Ross 308 broilers were collected from the slaughter line for an individual assessment. The collection was done after 9 days from February 15, 2018, to April 26, 2018, whereby on each day, between 10 and 32 samples were collected. An assessment scheme, revised according to the Welfare Quality assessment protocol for poultry (Welfare Quality®, 2009) and validated by Louton et al. (2020a), was used for the visual assessment ("macro score") of hock burn ranging from Macro Score 0 (no lesion), Macro Score 1 (superficial, attached (single) lesion or several single superficial or deep lesions ≤0.5 cm), Macro Score 2 (deep lesion >0.5 cm to ≤1 cm or superficial lesion >0.5 cm), Macro Score 3 (deep lesion >1.0 cm) to Macro Score 4 (whole hock extensively altered). A photographic documentation of each hock was performed using a Sony Cyber-shot DSC-RX100 digital camera (Sony Europe Limited, Surrey, UK).

Phase 2 (Validation) In the second part of the project, 500 hocks from Ross 308 broilers were collected from the slaughter line for an individual assessment (consisting of 133 samples of Macro Score 0, 95 of Macro Score 1, 173 of Macro Score 2, 96 of Macro Score 3, and 3 of Macro Score 4). The sample collection was done from July 26, 2019, to December 11th, 2019. After 7 days of sampling, between 35 and 120 samples were taken per day. The hocks were similarly assessed using the same macro scores as in the calibration phase, and a photograph of each hock was taken.

Assessment by Camera System

After the visual assessment, the photographic documentation and the measurement of the size of alterations, each hock was individually placed back into the evisceration line, and a picture of the alteration was taken by the camera system that was used by the slaughterhouse. Thus, all hocks were removed from the slaughter line for visual assessment and then placed back into the shackles for the assessment by camera. The camera used was a 1.3 MP color camera (IDS Imaging, Obersulm, Germany) with the "Chicken Check" software of CLK GmbH (Altenberge, Germany) and the Halcon software of MVTec Software GmbH (Munich, Germany) for image processing. With this procedure, the size of the alteration in percent (%) measured by the camera system was evaluated ("camera score") for each of the 200 (Phase 1) and 500 hocks (Phase 2). For each hock, the size of the surface of the hock was assessed by the camera and denoted as "reference surface area". For each picture taken by the automatic assessment system, 2 different software versions were used to evaluate possible optimizations in terms of the detection of alterations. These versions will be denoted "original" and "updated" software in the following. The updated camera score represents results from the updated software. For this update, the camera detection limits were adapted and the size of the surface of the hock ("reference surface area") assessed by the camera system was reduced while the threshold for the detection of differences in contrast was reduced. During the automatic camera assessment, the light intensities in lux (lx) (LMT Pocket-Lux 2B, LMT Lichtmesstechnik GmbH, Berlin, Germany), the relative humidity in percent (%) and the temperature in degrees Celsius (°C) (Testo 410-1 Flügelrad-Anemometer, Testo North America, West Chester, PA) were measured at the height of the hocks once per sampling session.

Phase 1 (Calibration) In the first phase, camera scores were allocated for each macro score. Threshold values of camera scores for respective macro scores were defined and the performance measures of the respective diagnosis evaluated.

Phase 2 (Validation) In the second phase, the threshold values that had been generated in the calibration phase were validated in 500 hocks. Furthermore, errors in the assessment of the reference surface areas and of lesions as well as possible impacts of climatic circumstances (temperature, humidity, and light intensity) were analyzed.

Statistical Methods

Multinomial logistic regression models for categorical data were used to predict the conditional probabilities of macroscopic findings given specific camera scores. Camera score cut-off values for macro score categories were derived by choosing the macro score categories with the highest probability according to the fitted models. To evaluate the cut-off values, performance measures for the classification were used as presented in Louton et al. (2020a). Values close to 1.0 were in favor.

Multinomial logistic regression models for categorical data were also used to measure the effect of software versions (original and updated) on error assessment for reference surface areas and lesions. Here, information of software versions was used as a predictor in the model, thus estimating the conditional probabilities (risk) of the different error types given the software version. Another multinomial logistic regression model for categorical data was used to analyze the relationship between the 2 error types. In this model, the conditional probabilities (risk) in the error assessment for reference surface areas were estimated given the error assessment for lesions.

Finally, further multinomial logistic regression models were used to measure the effects of temperature, humidity, and light intensity on error assessment for reference surface areas and lesions.

Results are presented as estimated risks and their corresponding 95% uncertainty intervals. Comparisons are presented by risk ratios, their corresponding 95% uncertainty intervals, and *P*-values.

RESULTS AND DISCUSSION

The results of the interobserver reliability were presented by Louton et al. (2020a). The PABAK value of interobserver reliability ranged from 0.75 to 0.87, with an average of 0.81.

Calibration of the Camera Scores, Thresholds Values

Initially, the distribution of camera scores for the original (Figure 1A) and for the updated camera score (Figure 1B) is presented descriptively for the macro scores (0−4). The updated camera score (Figure 1B) is based on a detection-adapted software update in which the size of the assessed hock surface was reduced and combined with a lowering of the threshold for the detection of differences in contrast. Both camera scores simulate 5 categories of the macro score (0−4). A rising camera score (original and updated software) was accompanied by a progressive increase in the macroscopic score, and most macro scores were easy to identify and distinguish, except for the second score (Macro Score 1), which overlapped partially with Macro Score 0. This increase in camera scores with a rising macro score was expected and observed by other authors (Louton et al., 2020b). In a recent study on the validation of foot pad dermatitis, similar results were observed (Louton et al., 2022). For the adjustment and calibration of camera scores, subsequently the highest probabilities of camera scores for given macro scores were estimated. When we used the updated software, the probabilities of camera scores indicating certain macro scores were slightly shifted to the right, thus resulting in higher camera scores for the respective macro scores. Table 1 depicts examples of “visual” (manually taken) pictures of hocks and the respective (automatically taken) pictures evaluated with the original and the updated camera score.

The camera score represents the percentage of the altered surface of a hock; it is a metrical value and can reach values of any positive number, including 0.00. For the detection of threshold values for the diagnosis of certain macro scores, the highest estimated probabilities of camera scores were used. The thresholds for the use of the original and updated camera scores are presented in Table 2. The presented thresholds show the camera score values at which the respective macro score had the highest probability of appearance. Macro Score 1 (small lesions up to 0.5 cm) had the highest probability at a camera score of ≥0.2 if the original camera score was used and ≥0.1 if the updated camera score was used. Macro Score 4 was most likely at a camera score of ≥6.8 (original software) or ≥11.4 (updated software). Note that these numbers represent the percentage and not permille of the altered surface, although permille is commonly used because of the large surface of the hock. In a comparable study on the detection of foot pad dermatitis, higher camera scores for the macro score 1 of foot pad lesions were observed (Louton et al., 2022). Thus, the results show that camera scores should be adapted to the specific kind of lesion, since the reference surface area of a hock is larger than a foot pad, smaller camera scores need to be applied. Automatic assessment systems at slaughter can be adjusted to these threshold values to classify the lesions. In the Netherlands, hock burn is

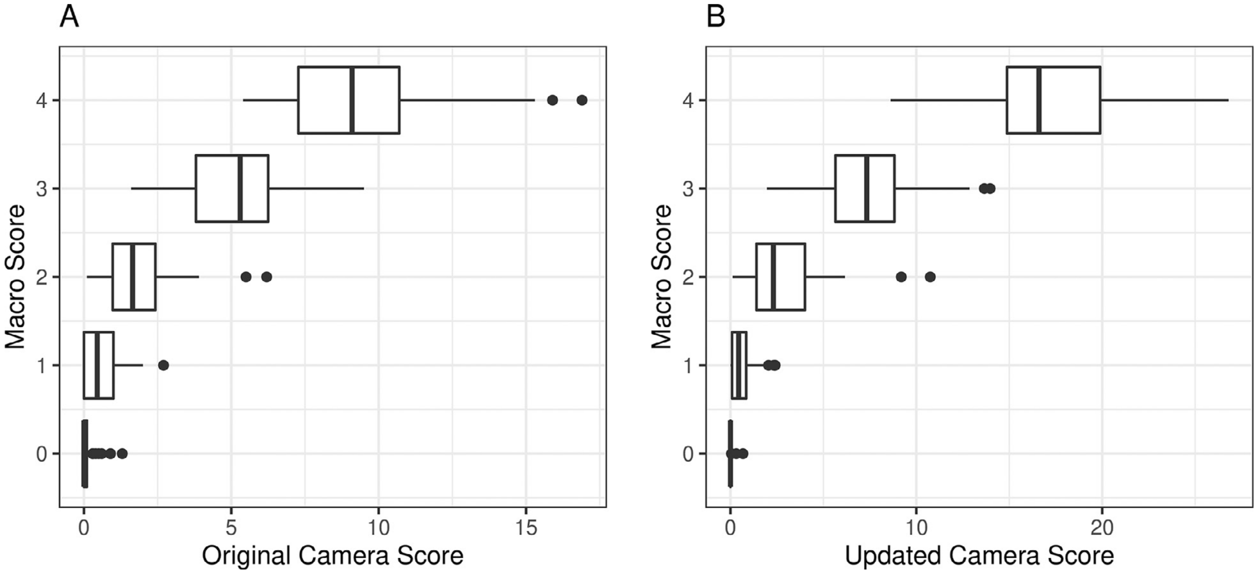
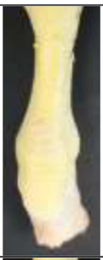
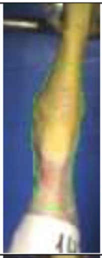
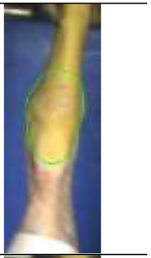
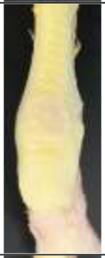
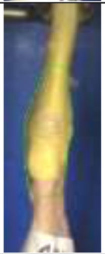
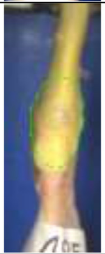
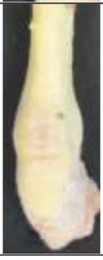
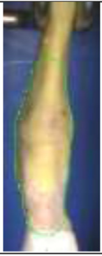
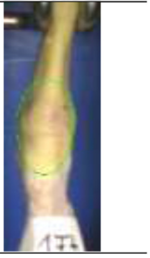


Figure 1. Descriptive presentation of the distribution of the camera scores (A: “original camera score”, B: “updated camera score”) at given macro scores (0−4) of hock burn lesions (according to Louton et al., 2020a) during the calibration phase ($n = 200$ hocks). Camera scores represent the percentage of the size of the alteration relative to the reference surface area of the hock.

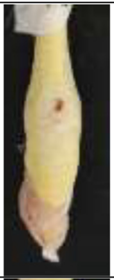
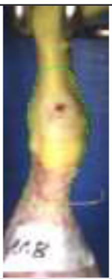
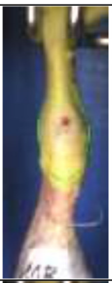
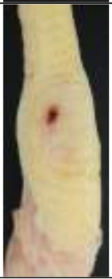
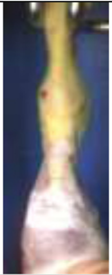
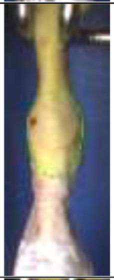
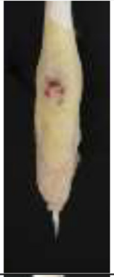
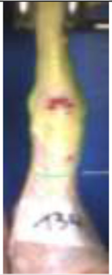
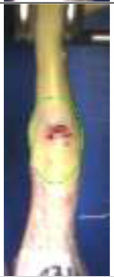
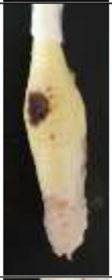
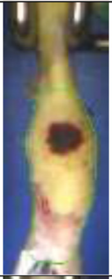
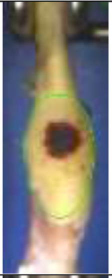
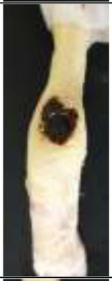
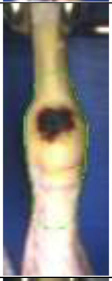
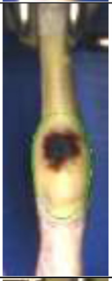
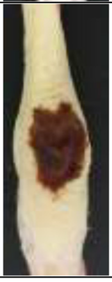
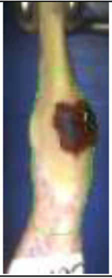
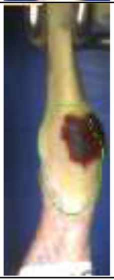
Table 1. Examples of assessed hocks with the visual and the respective camera scores, the latter based on the original or the updated software version.

Visual Score	Visual Picture	Camera Score, Original	Camera Picture, Original	Camera Score, Updated	Camera Picture, Updated
0		0.00		0.00	
0		0.00		0.00	
1		0.00		0.00	

(*continued*)

assessed at slaughter by camera systems that do not register the percentage of the lesions in relation to the hock; instead, the systems register the actual size of the hock burn lesion, and lesions are scored as hock burn if their size is above 0.5 cm (Pluimned, 2019). The adaptation of the presented software in the current study to this kind of assessment would be possible and could enable a standardized and consistent evaluation of hock burn in different member states of the European Union. The hock burn scoring system presented by Louton et al. (2020a) likewise classifies lesions in steps of 0.5 cm and could be used for validation. The assessment of hock burn lesions by size of the lesion and not by percentage would furthermore support the findings of Heitmann et al. (2020), who discussed the difficulty of evaluating the size of lesions of foot pad dermatitis by an observer using subjective measures. Observer bias and visual perception could influence the results, according to these authors. To ease the validation of automatic assessment systems, the scoring based on the size of hock burn lesions is advisable. Additionally, it would reduce the risk of errors in the assessment of the reference surface area because only the lesions but not the reference surface area is assessed by this system.

To evaluate the application of the defined threshold values, we used performance measures for the classification of criterions as presented by Louton et al. (2020a). Values close to 1.0 were in favor. The quality criterions and performance measures for the defined threshold values are presented in Table 3. The first 5 data columns depict the "confusion matrix," which represents the allocation of the actual health categories (macro scores, rows) to the defined threshold values of the attributed camera score (columns). The numbers represent the absolute numbers of the sample sizes. The last 4 data columns present the performance measures and quality criterions and the accuracy of the overall performance. The overall performance had an accuracy of 65% (original camera score) and 73% (updated camera score). Macro scores of 1 (small lesions up to a size of 0.5 cm) were diagnosed with a sensitivity of only 0.32 (original camera score) or 0.48 (updated camera score). However, considering the other performance measures and the confusion matrix, one should note that the other macro scores (0, 2, 3, and 4) were diagnosed with a sensitivity of 0.62 to 0.95 if the above-stated threshold values (see Table 2) were used. Thus, the evaluated automatic assessment system identified larger lesions (i.e., those above 0.5 cm) more easily and more reliably, with a higher sensitivity and specificity, than small lesions. This is in accordance to results on a study of the validation of automatic assessment of foot pad dermatitis

1		0.50		0.80	
1		0.60		0.54	
2		2.40		4.00	
3		9.50		13.97	
4		10.90		18.18	
4		15.30		23.07	

Green lines represent the reference surface area of the hock, red lines the identified lesions of hock burn.

Table 2. Threshold values (minimum, maximum) of the camera scores (based on the original or updated software version) for the macro scores (0−4) of hock burn lesions (according to Louton et al., 2020a) during the calibration phase ($n = 200$ hocks).

Macro score	Minimum	Maximum
	Original camera score	
0	0.00	0.19
1	0.20	0.69
2	0.70	3.01
3	3.02	6.74
4	6.75	Inf
	Updated camera score	
0	0.00	0.10
1	0.11	1.05
2	1.06	4.36
3	4.37	11.42
4	11.43	Inf

Abbreviation: Inf, infinite.

(Louton et al., 2022). However, the result contrasts with the findings by Vanderhasselt et al. (2013), who evaluated an automatic assessment system for foot pad dermatitis in broilers. These authors observed that foot pads that were not scored by the camera system as being altered were more likely to be rated with a high score by the experts.

Validation of Threshold Values

The threshold values that had been determined in the calibration phase were validated in a second step (Table 4). With these threshold values and the original camera score, solely Macro Score 0 was diagnosed with satisfying performance measures. With the threshold values and the updated camera score, Macro Scores 0, 2, 3, and 4 were identified with a sensitivity of 0.55 to 0.96 and an accuracy of 0.60. The lower performance measures of high macroscopic scores in this phase in our study might be explained by a smaller number of hocks with high macroscopic scores and thus a small sample size, especially of Macro Score 4 ($n = 3$). Jung et al. (2021) reached a high sensitivity (0.95) and a specificity of 0.77 in their study considering the automatic assessment of keel bone deviations in laying hens after several optimization phases. In their study, the sensitivity varied from 0.28 to 0.95 and the specificity from 0.60 to 0.90. Authors who used automatic assessment systems to evaluate animal feed intake on farm by sound recordings reported that 86% of the feed intake was correctly monitored (Aydin et al., 2015). In another study, the authors reported that the correlation of the number of peckings and feed intake was high and concluded that a sound detection system has potential as a tool to monitor feed intake (Aydin et al., 2014). Zhuang et al. (2018) even reached an accuracy of above 99% when evaluating the health of broilers on farm. We agree with other authors (Vanderhasselt et al., 2013; Louton et al., 2020b; Jung et al., 2021) that automatic assessment systems are promising tools to evaluate welfare conditions in livestock and can be used to reflect the health of animals. However, threshold values need to be chosen carefully and should be validated to provide reliable results. This would be necessary for the respectively used camera and software system, the conditions at slaughter and potentially for the assessed birds type (size of the hocks).

Errors of Assessments of Reference Surface Areas and Lesions

Figure 2 depicts the risk of possibly occurring errors during the assessment of the reference surface area by the camera system for the original and updated software

Table 3. Performance measures of predicted macro scores at given camera scores (for the original and the updated camera score) for the evaluation of hock burn lesions in broilers during the calibration phase of the camera system ($n = 200$ hocks).

	Prediction ($n = 200$ hocks)					Performance			
			Original camera score						
	<0.20	≥0.20 to <0.70	≥0.70 to <3.02	≥3.02 to <6.75	≥6.75		Accuracy = 0.65		
Macro score	0	1	2	3	4	Sens	Spec	PPV	NPV
0	31	6	3	0	0	0.78	0.91	0.69	0.94
1	13	13	14	0	0	0.32	0.92	0.50	0.84
2	1	7	25	7	0	0.62	0.87	0.54	0.90
3	0	0	4	28	8	0.70	0.91	0.67	0.92
4	0	0	0	7	33	0.82	0.95	0.80	0.96
			Updated camera score						
	<0.11	≥0.11 to <1.06	≥1.06 to <4.37	≥4.37 to <11.43	≥11.43		Accuracy = 0.73		
Macro score	0	1	2	3	4	Sens	Spec	PPV	NPV
0	38	2	0	0	0	0.95	0.92	0.76	0.99
1	12	19	9	0	0	0.48	0.96	0.73	0.88
2	0	5	27	8	0	0.68	0.91	0.64	0.92
3	0	0	6	27	7	0.68	0.92	0.67	0.92
4	0	0	0	5	35	0.88	0.96	0.83	0.97

The cut-off value was set at specifically evaluated threshold camera scores.
Abbreviations: NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity, Spec, specificity.

Table 4. Performance measures of predicted macro scores at given camera scores (for the original and the updated camera score) for the evaluation of hock burn lesions in broilers during the validation phase of the camera system ($n = 500$ hocks).

	Prediction ($n = 500$ hocks)					Performance			
	Original camera score								
	<0.20	≥0.20 to <0.70	≥0.70 to <3.02	≥3.02 to <6.75	≥6.75	Accuracy = 0.65			
Macro score	0	1	2	3	4	Sens	Spec	PPV	NPV
0	126	4	3	0	0	0.95	0.75	0.58	0.98
1	70	14	10	1	0	0.15	0.92	0.29	0.82
2	19	24	100	30	0	0.58	0.86	0.68	0.79
3	2	6	32	48	8	0.50	0.92	0.60	0.89
4	0	0	2	1	0	0.00	0.98	0.00	0.99
	Updated camera score								
	<0.11	≥0.11 to <1.06	≥1.06 to <4.37	≥4.37 to <11.43	≥11.43	Accuracy = 0.60			
Macro score	0	1	2	3	4	Sens	Spec	PPV	NPV
0	128	4	1	0	0	0.96	0.78	0.61	0.98
1	65	21	8	1	0	0.22	0.89	0.32	0.83
2	13	32	98	30	0	0.57	0.88	0.71	0.79
3	4	9	30	53	0	0.55	0.92	0.63	0.90
4	0	0	1	0	2	0.67	1.00	1.00	1.00

The cut-off value was previously set at specifically evaluated threshold camera scores during the calibration phase.
Abbreviations: NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity, Spec, specificity.

during the validation phase. The probability of a correct detection of the reference surface area was higher if the original software was used (risk = 0.69 [0.65−0.75]) than with the use of the updated software (risk = 0.36 [0.31−0.41]). This difference in risk for the correct identification of reference surface areas between the software versions was significant (relative risk [**RR**] original vs. updated software = 1.933 [1.671−2.286]; $P < 0.001$). The risk for a faulty detection of the reference surface area as being too large was significantly higher with the original software (risk = 0.29 [0.24−0.33]) than with the updated software (risk = 0.01 [0.0−0.02]; RR original vs. updated software = 27.874 [12.924−75.659]; $P < 0.001$). The risk for a "too small" identification of the reference surface area was significantly higher with the updated software (risk = 0.62 [0.58−0.67]) than with the original software (risk = 0.02 [0.01−0.03]; RR original vs. updated software = −0.605 [0.66 to −0.556]; $P < 0.001$). The risk of a completely wrong or shifted detection of the reference surface area was nearly 0 in both software versions. Especially valgus or varus deformations can lead to an angular deviation of the tibiotarsal-tarsometatarsal joint (Leterrier and Nys, 1992). Van den Brand et al. (2022) stated an incidence of 1.8 to 66% of affected broilers with large variations within flocks. Such angular deviations could lead to a tilting of the hock and thus to a completely wrong or shifted detection of the reference surface area as seen in the Table 1 in the last hock.

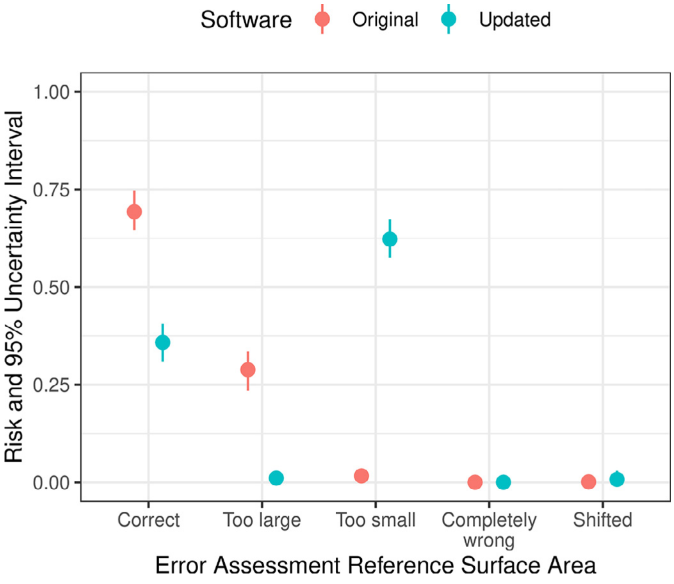


Figure 2. Estimated probabilities of occurrence of errors in the camera-based assessment of reference surface area of the hock with the original and the updated camera software (n = 500 hocks; validation phase).

We also evaluated the risk of a faulty detection of lesions of the hock by the camera system (Figure 3). The hock burn lesions either were identified correctly or were detected as too large, too small, or not at all. Both software versions had the highest probability of identifying the lesions correctly. The probability of a correct identification of lesions by the camera system was slightly higher with the updated software (risk = 0.66 [0.62 −0.70]) than with the original software (risk = 0.63 [0.58−0.67]), although the difference was not significant (RR original vs. updated software = 0.948 [0.865 −1.040]; $P = 0.119$). The risk of a too large identification of lesions was 2 times higher with the original software (risk = 0.10 [0.07−0.13]) than with the updated software (risk = 0.05 [0.03−0.07]), although the difference was not significant (RR original vs. updated software = 1.963 [1.253−3.187]; $P = 0.998$). The risk of a too small identification of the hock burn lesions did not differ between the original (risk = 0.13 [0.10−0.16]) and the updated

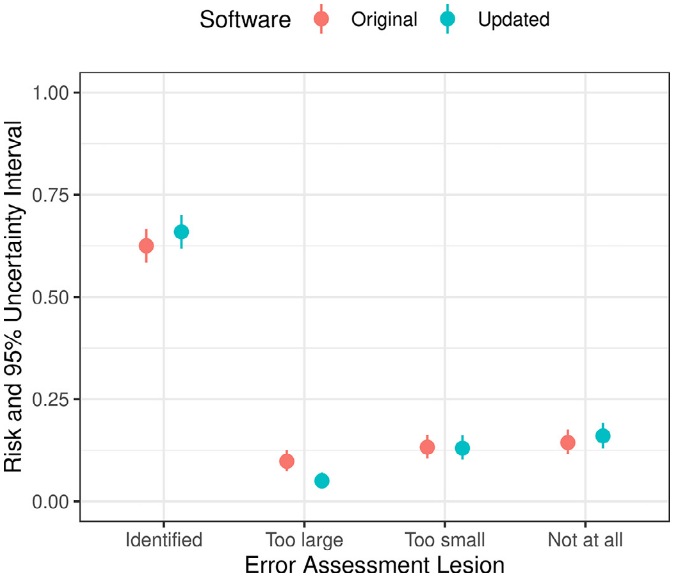


Figure 3. Estimated probabilities of occurrence of errors in the camera-based assessment of hock burn lesions with the original and the updated camera software ($n = 500$ hocks; validation phase).

(risk = 0.13 [0.10–0.16]) software (RR original vs. updated software = 0.002 [−0.039 to 0.045]; $P = 0.455$). Furthermore, the risk of lesions not being detected at all did not differ between the software versions (risk original software = 0.14 [0.12–0.18]; risk updated software = 0.16 [0.13–0.19]; RR original vs. updated software = −0.016 [−0.060 to 0.029]; $P = 0.232$).

In a further analysis, we assessed a potential association between the faulty detection of the reference surface areas and possible errors in the assessments of lesions. Using the original (Figure 4) and the updated (Figure 5) software, the probability of correctly identifying lesions was significantly higher than the probability of a faulty detection (too large, too small, not at all) if the reference surface area was identified correctly. If the reference surface area was detected as too large, the probability of a correct identification of lesions was significantly higher than the risk of a faulty detection of lesions if the original software was used. Furthermore, with the original software, the risk of the “not at all” detection of lesions was significantly higher if the reference surface area was shifted than if reference surface areas were identified as too small or too large. Using the updated software, the probability of a correct identification of lesions was significantly higher than the risk of a faulty detection of lesions if the reference surface area was identified as too small. If the reference surface area was identified as too large with the updated software, the risk of identifying lesions as too large was significantly higher than the risk of identifying them as too small or as shifted. The other observed faults in the detection of the reference surface area were not associated with mistakes in the detection of lesions, neither with the original nor with the updated software. These results and the low rate of errors are very promising. Other authors reported that an automatic assessment system for foot pad dermatitis in broilers at slaughter indicated a lesion in 49.5% of broilers where a lesion was not present (Vanderhasselt et al., 2013). For the assessment system used in our study, we can conclude that even if the original software assessed the reference surface area as too large or the updated software assessed the reference surface area as too small, a correct rather than a faulty identification of lesions had the highest probability.

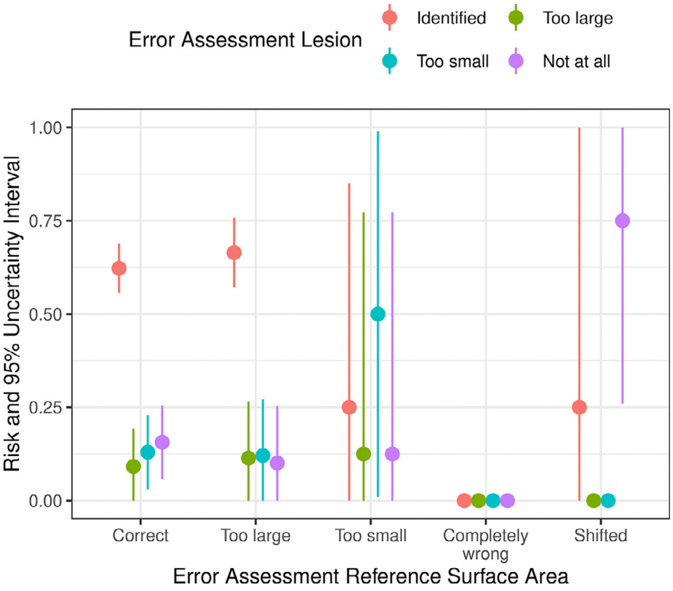


Figure 4. Estimated risk and uncertainty interval (95%) of the occurrence of errors during the assessment of hock burn lesions (identified, too large, too small, not at all) by the camera system with the original software and the association to the faulty detection of the reference surface area of the hock (correct, too large, too small, completely wrong, shifted; $n = 500$ hocks; validation phase).

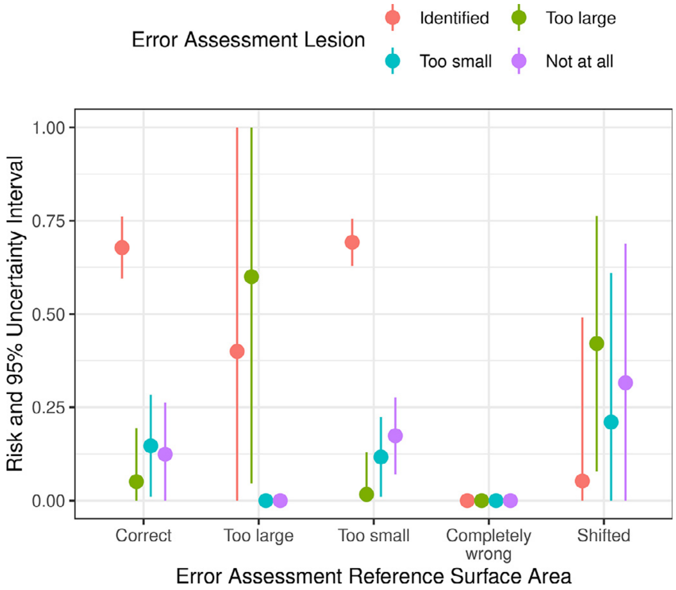


Figure 5. Estimated risk and uncertainty interval (95%) of the occurrence of errors during the assessment of hock burn lesions (identified, too large, too small, not at all) by the camera system with the updated software and the association to the faulty detection of the reference surface area of the hock (correct, too large, too small, completely wrong, shifted; $n = 500$ hocks; validation phase).

Effects of Temperature, Humidity, and Light Intensity on Errors

The climatic conditions at slaughter demand a high resistance of camera and software systems toward temperature, humidity, and light intensities; the latter, for example, could be too dimmed or cause an overexposure of light. To evaluate the susceptibility of the assessment system toward temperature, humidity, and light intensities, we analyzed possible errors that occurred in relation to the climatic circumstances in a multinomial regression model. It was not possible to predict the risk ratios with sufficient certainty; therefore, separate evaluations were done. The Supplementary Material (Figures S1 and S3) depicts the possible correct and false detection of the reference surface area (correct, too large, too small, shifted) with the original and the updated software depending on the temperature (in degrees Celsius) and the estimated risk and uncertainty interval (95%) of the occurrence of errors depending on the temperature (in degrees Celsius). We found that the correct detection of the reference surface area by the original software increased with rising temperatures, whereas the results of the updated software were not affected by temperature variations. A higher humidity increased the probability of a correct detection of reference surface areas, without differences between the software versions (Supplementary Material, Figures S1 and S4). It is expected that temperature and humidity at slaughter could affect the technology of automatic assessment, especially because of condensation and fogging of the camera. The technical devices at slaughter experience severe humidity. Considering the light intensity, the correct detection of reference surface areas was affected if the original software was used and decreased with increasing light intensities (Supplementary Material, Figures S1 and S5). Especially an overexposure of light should be regarded as cause for the decreasing probability of correct identification of reference surface areas when using the original software.

Considering the faulty identification of lesions, we observed only minor effects of temperature, humidity, and light intensity (Supplementary Material, Figure S2). Likewise, even though only minor effects were observed, tendencies of decreasing correct identification of hock burn lesions with increasing humidity and increasing light intensities are demonstrated, whereas the temperature seemed to have no effect (Supplementary Material, Figures S6−S8). It seems that the updated software version is not as susceptible toward climatic factors, especially toward light intensity.

CONCLUSIONS

With the presented threshold values, the 5 macro scores of an adapted scheme (a modification of the Welfare Quality assessment protocol for poultry) for the assessment of hock burn in broilers can be used to classify hock burn by an automatic assessment system. Automatic assessment systems are promising tools to evaluate welfare conditions in livestock and can be used to reflect the health of animals. However, threshold values need to be chosen carefully and should be validated to provide reliable results. Software adaptations can be used to improve the performance measures of diagnosis and to lower the probability of errors during automatic assessment.

ACKNOWLEDGMENTS

The authors thank Dr. Josef Bachmeier for the support and organization of the study. We thank Dr. Verena Lietze for the scientific language editing. We thank Stefan and Yeliz Wiesbeck for their valuable and dedicated support during data collection and Dr. Hanka Sanftleben for assistance in graphical visualization.

This project was supported by funds from of the Federal Ministry of Food and Agriculture (BMEL) based on a decision from the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program (grant number: FKZ 2817903715).

Ethical statement: Any aspect of the work covered in this manuscript that has involved animals has been conducted according to the international and national guidelines for humane animal treatment, in compliance with relevant legislation and with the high ethical standards concerning animal welfare of the IAVE (International Association of Veterinary Editors) Guidelines. All animals used were housed according to the German Order on the Protection of Animals and the Keeping of Production Animals prior to slaughter. Sampling procedures were in accordance with the German Animal Welfare Act, the regulations on the welfare of animals used for experiments or for other scientific purposes, and national and international regulations on the welfare of animals at slaughter (Council Regulation [EC] No 1099/2009, of 24 September 2009, on the protection of animals at the time of killing). Ethical review and approval were waived for this study because feet of the broilers (samples for the research) were taken after slaughter from dead animals that were primarily rejected for use in the food chain for other reasons. All mandatory laboratory health and safety procedures were complied with during sampling procedures.

DISCLOSURES

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.psj.2022.102025.

REFERENCES

Aydin, A. 2017. Using 3D vision camera system to automatically assess the level of inactivity in broiler chickens. Comput. Electron. Agric. 135:4–10.

Aydin, A., C. Bahr, and D. Berckmans. 2015. A real-time monitoring tool to automatically measure the feed intakes of multiple broiler chickens by sound analysis. Comput. Electron. Agric. 114:1–6.

Aydin, A., C. Bahr, S. Viazzi, V. Exadaktylos, J. Buyse, and D. Berckmans. 2014. A novel method to automatically measure the feed intake of broiler chickens by sound technology. Comput. Electron. Agric. 101:17–23.

Ben Sassi, N., X. Averós, and I. Estevez. 2016. Technology and poultry welfare. Animals 6:62.

Bergmann, S., H. Louton, C. Westermaier, K. Wilutzky, A. Bender, J. Bachmeier, M. Erhard, and E. Rauch. 2016. Field trial on animal-based measures for animal welfare in slow growing broilers reared under an alternative concept suitable for the German market. Berl. Munch. Tierarztl. Wochenschr. 129:453–461.

Bessei, W. 2006. Welfare of broilers: a review. Worlds Poult. Sci. J. 62:455–466.

Butterworth, A. 2013. On-farm broiler welfare assessment and associated training. Braz. J. Poult. Sci. 15:71–77.

Byrt, T., J. Bishop, and J. B. Carlin. 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46:423–429.

Dawkins, M. S., S. J. Roberts, R. J. Cain, T. Nickson, and C. A. Donnelly. 2017. Early warning of footpad dermatitis and hockburn in broiler chicken flocks using optical flow, bodyweight and water consumption. Vet. Rec. 180:499.

European Union. 2016. Use of slaughterhouse data to monitor welfare of broilers on farm. Overview report of a series of audits of the Directorate-General for Health and Food Safety from 2013 to 2015 to evaluate the official controls on the welfare of chickens kept for meat production using slaughterhouse data to establish farm checks. European Commission. Accessed April 2022. https://op.europa.eu/en/publication-detail/-/publication/9fbf913d-de15-11e6-ad7c-01aa75ed71a1/language-en.

Greene, J. A., R. M. McCracken, and R. T. Evans. 1985. A contact dermatitis of broilers—clinical and pathological findings. Avian Pathol. 14:23–38.

Haslam, S. M., T. G. Knowles, S. N. Brown, L. J. Wilkins, S. C. Kestin, P. D. Warriss, and C. J. Nicol. 2007. Factors affecting the prevalence of foot pad dermatitis, hock burn and breast burn in broiler chicken. Br. Poult. Sci. 48:264–275.

Heitmann, S., J. Stracke, H. Petersen, B. Spindler, and N. Kemper. 2020. First approach validating a scoring system for foot-pad dermatitis in broiler chickens developed for application in practice. Prev. Vet. Med. 154:63–70.

Hepworth, P. J., A. V. Nefedov, I. B. Muchnik, and K. L. Morgan. 2011. Hock burn: an indicator of broiler flock health. Vet. Rec. 168:303.

Johnsen, P. F., T. Johannesson, and P. Sandøe. 2001. Assessment of farm animal welfare at herd level: many goals, many methods. Acta. Agric. Scand. A. Anim. Sci. 51:26–33.

Jung, L., A. Nasirahmadi, J. Schulte-Landwehr, and U. Knierim. 2021. Automatic assessment of keel bone damage in laying hens at the slaughter line. Animals 11:163.

Kjaer, J. B., G. Su, B. L. Nielsen, and P. Sørensen. 2006. Foot pad dermatitis and hock burn in broiler chickens and degree of inheritance. Poult. Sci. 85:1342–1348.

Leterrier, C., and Y. Nys. 1992. Clinical and anatomical differences in varus and valgus deformities of chick limbs suggest different aetio-pathogenesis. Av. Pathol. 21:429–442.

Louton, H., S. Bergmann, A. Piller, M. Erhard, J. Stracke, B. Spindler, P. Schmidt, J. Schulte-Landwehr, and A. Schwarzer. 2022. Automatic scoring system for monitoring foot pad dermatitis in broilers. Agriculture 12:221.

Louton, H., S. Bergmann, S. Reese, M. Erhard, J. Bachmeier, B. Rösler, and E. Rauch. 2018b. Animal- and management-based welfare indicators for a conventional broiler strain in two barn types (Louisiana barn and closed barn). Poult. Sci. 97:2754–2767.

Louton, H., M. Erhard, K. Wirsch, S. Bergmann, A. Piller, P. Schmidt, and E. Rauch. 2020b. Comparison of four assessment methods of foot pad dermatitis and hock burn of broilers. Berl. Munch. Tierarztl. Wochenschr. 133:1–11.

Louton, H., M. Erhard, and A. C. Wöhr. 2018a. Acquisition of animal-based welfare measures at slaughter of poultry. Fleischwirtschaft 98:94–98.

Louton, H., A. Piller, S. Bergmann, M. Erhard, J. Stracke, B. Spindler, N. Kemper, P. Schmidt, B. Schade, B. Boehm, E. Kappe, J. Bachmeier, and A. Schwarzer. 2020a. Histologically validated scoring system for the assessment of hock burn in broilers. Avian Pathol. 49:230–242.

Maisack, C. 2016. Verpflichtung zur Erhebung tierbasierter Indikatoren am Schlachthof nach dem EU und dem deutschen Recht. Pages 71−73 in Internationale Gesellschaft für Nutztierhaltung. im Fokus Nutztierhaltung, ed. Tierschutzindikatoren am Schlachthof, Munich, Germany.

McKeegan, D. 2010. Foot pad dermatitis and hock burn in broilers: risk factors, aetiology and welfare consequences. Faculty of Veterinary Medicine, University of Glasgow, Scotland, UK Research project final report.

Pluimned. 2019. Bijlage 8 Beoordelingssysteem vleeskuikens. Version 5. https://www.avined.nl/wp-content/uploads/8-Beoordelingssysteem-vleeskuikens-IKB-Kip-versie-5-190601.pdf (Accessed June 2021).

Saraiva, S., C. Saraiva, and G. Stilwell. 2016. Feather conditions and clinical scores as indicators of broiler welfare at the slaughterhouse. Res. Vet. Sci. 107:75–79.

van den Brand, H., R. Molenaar, and M. Klaasen. 2022. Research note: comparing methods to assess Valgus-Varus deformity in broiler chickens. Poult. Sci. 101:101907.

Vanderhasselt, R. F., M. Sprenger, L. Duchateau, and F. A. M. Tuyttens. 2013. Automated assessment of footpad dermatitis in broiler chickens at the slaughter-line: evaluation and correspondence with human expert scores. Poult. Sci. 92:12–18.

Welfare Quality® Network. 2018. Towards a Welfare Quality® assessment system. Fact sheets. http://www.welfarequalitynetwork.net/media/1043/fact_sheet_towards_a_welfare_quality_assessment_system_-_english.pdf (Accessed May 2021).

Welfare Quality®. 2009. Welfare Quality® assessment protocol for poultry (broiler and laying hens). Presented on October 9th, 2009, at the Animal Welfare Conference in Uppsala, Sweden ISBN: 978-90-78240-06-8.

Zhuang, X., M. Bi, J. Guo, S. Wu, and T. Zhang. 2018. Development of an early warning algorithm to detect sick broilers. Comput. Electron. Agric. 144:102–113.