Improving intra- and inter-observer repeatability and accuracy of keel bone assessment by training with radiographs

Sabine G. Gebhardt-Henrich,[1] Christina Rufener,[2] and Ariane Stratmann

Center for Proper Housing: Poultry and Rabbits (ZTHZ), Division of Animal Welfare, VPH Institute, University of Bern, Zollikofen, 3052, Switzerland

ABSTRACT Assessing keel bone damage reliably and accurately is a requirement for all research on this topic. Most commonly, assessment is done on live birds by palpation and is therefore prone to bias. A 2-day Training School of the COST Action "Identifying causes and solutions of keel bone damage in laying hens" with 16 participants of variable experience was held where palpation of live hens was followed by consulting corresponding radiographic images of keel bones. We hypothesized that the inter-observer and intra-observer repeatabilities as well as the agreement between palpation and assessment from the radiograph (considered as the accuracy) would increase from day 1 to 2. Repeatability estimates were calculated using the R-package rptR and the change in level of accuracy on day 1 and 2 was analyzed with generalized linear models. As predicted, the inter-observer repeatabilities of the assessments of the fractures and deviations were improved by training, but this improvement differed for fractures and deviations between the cranial, middle, and caudal parts of the keel bone. Intra-observer repeatabilities before training also differed between the different parts of the keel bone and were highest for fractures at the caudal part of the keel bone. The training affected the accuracy of palpation to different degrees for the different parts of the keel bone. A training effect was found for the caudal part of the keel bone in regard to fractures and deviations, but for fractures the training effect was missing for the cranial part and for deviations it was missing for the middle part of the keel bone. In conclusion, the training school involving radiographs improved inter-observer repeatabilities in the diagnosis of fractures and deviations of keel bones and thus had the potential to lead to more comparable results among research groups.

Key words: laying hen, keel bone, palpation, radiograph, repeatability

2019 Poultry Science 98:5234–5240
http://dx.doi.org/10.3382/ps/pez410

INTRODUCTION

A majority of laying hens damage their keel bone during their lifetime which includes fractures and deviations of various severity at different locations of the bone constituting a severe welfare problem (Riber et al., 2018). Keel bone damage has been found to various degree in different countries (Käppeli, et al., 2011b; Bestman and Wagenaar, 2014; Riber et al., 2018; Rørvang et al., 2019; Scholz et al., 2009), different genetic stock (Käppeli, et al., 2011a; Regmi et al., 2016; Stratmann et al., 2016; Heerkens, et al., 2016b; Eusemann et al., 2018), and different housing systems (Rodenburg et al., 2008; Wilkins et al., 2011; Hester et al., 2013; Petrik et al., 2015; Heerkens, et al., 2016a; Casey-Trott et al., 2017).

Unfortunately, comparisons across these studies are difficult. Researchers from different groups may assess the same kind of damage differently, and differentiating between the type of damage (i.e., fracture and/or deviation) as well as assessing severity is difficult. Inconsistent keel bone damage assessment is seen as a major problem inhibiting research on the causes and prevention of keel bone damage (Casey-Trott et al., 2015), and this problem is especially important for the most common diagnostic method, palpation of live hens. Unlike more exact methods such as radiography, peripheral quantitative computed tomography, or dissections of dead birds ideally followed by histology which are time- and cost-consuming, palpation can be easily and quickly done on a large number of live laying hens inside the barn. Therefore, standardizing the palpation method and the way of reporting keel bone damage is imperative and must be improved (Harlander-Matauschek et al., 2015).

Several studies addressed inter- and intra-repeatability of assessment of fractures and deviation in keel bones. Experience in palpating keel bones has been shown to improve both accuracy and agreement

Received March 25, 2019.
Accepted June 20, 2019.
[1]Corresponding author: sabine.gebhardt@vetsuisse.unibe.ch
[2]Present address: Department of Animal Science, University of California, Davis, One Shields Avenue, 95616 Davis, USA.

between assessors (Petrik et al., 2013) but its effect is limited and can even lead to lower specificity and precision (Buijs et al., 2018). Training has also been discussed as improving accuracy and agreement between assessors (Petrik et al., 2013; Buijs et al., 2018; Chargo et al., 2018). While training involved visualization of the keel bone by constructing 3 D models based on computed tomography (CT) in Chargo et al. (2018), most training sessions consisted in the simultaneous palpation of one or more experienced and several non-experienced assessors with discussions of the results (Petrik et al., 2013; Buijs et al., 2018). Additionally, to experience and training, the location of the fracture on the keel bone affects the accuracy of palpation (Buijs et al., 2018).

In order to improve consistency in palpations between research groups, a 2-day Training School of the COST Action CA15224 "Identifying causes and solutions of keel bone damage in laying hens" (http://www.keelbonedamage.eu/, accessed 22-10-2018) was held where 16 trainees with various levels of experience in the palpation of keel bones palpated live laying hens before and after consulting the radiographic image of the keel bone of the same hen. The goal was that the participants were able to fine-tune their palpation technique. We predicted that the inter- and intra-repeatability and accuracy of palpation (i.e., the agreement between the palpation and the "true" state as seen on the radiograph) would be higher on the second day than on the first day of training.

The aim of this study was to evaluate the effectiveness of using palpations along with the corresponding radiographs of keel bones in order to increase repeatability and accuracy of keel bone damage assessment in live hens.

MATERIALS AND METHODS

Ethical Statement

The use of animals was approved by the Cantonal license BE 74/17. All human participants gave their written consent for the analyses of their palpation assessments.

Animals

On the first day of the training school 40 Lohmann Selected Leghorn and 40 Lohmann Brown laying hens of 63 weeks of age representing a wide variety of keel bone damage, i.e., including minor and major damage at different localizations of the keel bone, were collected from a pen equipped with an aviary system (Bolegg Terrace, Krieger AG, Ruswil, Switzerland) and radiographed following Rufener et al. (2018). Briefly, hens were carefully hung upside down and a latero-lateral image was produced with a mobile radiograph unit (GIERTH HF 200ML; radiograph tube Toshiba D-124 with maximal acceleration voltage of 100 kV; radiograph plate Canon CXDI-50 G; software Canon CXDI Control Software NE) using a distance of 80 cm and voltage of 46 kV/2.4 mAs. The radiographs were performed about 4 h before the palpations and hens were kept in the wintergarden of the barn where further damage to the keel bone was unlikely until palpation. Just before palpations, the hens were placed into 4 different crates with 10 hens each. Four stations consisting of a crate situated next to a computer screen where the radiographic images of the corresponding hens were available were set up. Care was taken to include a great variety of keel bone damage for each station. Participants were distributed among the stations but some participants palpated hens from more than one station. A second set of another 40 hens was selected and radiographed according to the same protocol and displayed at the 4 stations on the second day. All hens could be identified by numbered legbands.

Assessors

From the list of applicants, 16 people from different groups (mostly from academic research groups, but also from companies using palpations of keel bones) were chosen. The aim was that these people would train their colleagues so participants came from different countries and had not palpated keel bones together before the training school. Some had palpated thousands of laying hens before, others had not done it at all and the experience of the rest was between these extremes.

Experimental Design

On the first and second training day, 16 and 15 participants, respectively (3 or 4 per station), conducted palpations of keel bones in the following way.

Session 1: Participants palpated 10 to 20 hens (from 1 to 2 stations) and noted whether they found fractures or deviations in the cranial third, the central third, or the caudal third of the keel bone. On the data sheets, a crude sideway image was provided where participants had tick marks to indicate whether they thought there was a fracture or deformation. The sideway image of the keel bone was divided into 3 equally sized parts. Thus, 6 binomial variables (i.e., caudal deviation, caudal fracture, middle deviation, middle fracture, cranial deviation, and cranial fracture) were recorded per hen.

Session 2: Participants palpated the same individual hens immediately after the first session and recorded their assessment in the same way in order to determine intra-observer reliability prior to training. Participants did not communicate with each other during the first 2 sessions of palpation.

Session 3: After writing down the results from the second session of palpation, participants looked at the radiograph of the hens and recorded a third, radiograph-based assessment. Participants were encouraged to discuss the results of the palpation and radiograph assessments with each other. Again, the same 6

binomial variables as described above were recorded per hen.

Sessions 1, 2, and 3 were repeated on a second set of hens in the same way on the following day. When participants made ambiguous notes their data were considered missing. Therefore, the number of participants varies in the analyses.

Statistical Analyses

In this paper, we follow the terminology of Stoffel et al. (2017) who use the term repeatability as a quantitative measure of reliability (sensu reliability; Bartlett and Frost, 2008). Inter- and intra-observer repeatability estimates based on the intra-class correlation coefficients were calculated using R (version 3.4.3), package "rptR" (Nakagawa and Schielzeth, 2010). The special feature of this package is that it can deal with binomial data and controls for fixed effects (Stoffel et al., 2017). The presence or absence of a fracture or deviation in the respective third of the keel bone (cranial, middle, or caudal) was taken as a binary variable, resulting in 6 binary variables (i.e., caudal deviation, caudal fracture, middle deviation, middle fracture, cranial deviation, and cranial fracture). The variance in the repeatability estimates was assessed by parametric bootstrapping meaning that the original model and sampling design were used to simulate a new data set and analyze it 1,000 times within this package (Stoffel et al., 2017). The model included the identity of the assessor and thus allowed for different repeatabilities among assessors.

Link-scale instead of original scale approximations of the variables are presented in this manuscript because they represent the assumed underlying scale and not the binary scale of the data (Schielzeth and Bolund, 2010). That means that the binary variable (e.g., fracture yes or no) is (roughly) replaced by the probability of being a yes or no. Repeatabilities were calculated for the 3 sections of the keel bone separately as well as pooling the results from the 3 locations. In the latter case, a keel bone was counted as fractured/deviated if a fracture/deviation was found in at least 1 location of the bone because most existing studies on keel bones report damage for the entire bone.

Inter-observer reliabilities were based on results from the first palpations (session 1) and before looking at radiographs among observers on either day. To test intra-observer repeatabilities for individual assessors, another variable was created (binary variable: fracture/deviation outcome of the first palpation session equals/not equals the fracture/deviation outcome of the second palpation session). Intra-observer reliabilities were based on the results from the first and second palpation (session 1 vs. session 2) within each observer of day 1 only. A Fisher's exact test (Proc Freq, SAS 9.4) on this variable was conducted.

For the hypothesis that the accuracy of palpation, i.e., the agreement between palpation and radiograph of the same hen was higher after training, 2 analyses were performed. First, R based on the package "rptR" was calculated by comparing the palpation before looking at the radiograph with the assessment of the radiograph for each assessor (= "accuracy"). Secondly, a generalized linear model was performed with Proc Glimmix SAS 9.4. following Bartlett and Frost (2008) by defining accuracy as the percentage of achieving the same state in a binomial variable in order to estimate the factors day, section of the keel bone, and assessor. The binary outcome variable was whether the assessment of individual assessors from day 1 was identical to the assessment from day 2 or not. The 2 assessments were the palpation in session 2 (= second palpation of the hen before discussing with other trainees and before seeing the radiograph) and the assessment of damage from the radiograph by the same trainee. The trainees could discuss the radiograph with the radiologist and other people that were present (other trainees at their station).The fixed effects were time point (day 1 or 2), the section of the keel bone (cranial, middle, caudal), and the assessor. All possible interactions (up to the 3-way interaction) were included at first but consecutively removed unless $P < 0.2$. A significant interaction was partitioned to yield tests of simple effects (Winer, 1971 in SAS Users' Guide, 2010). Multiple a posteriori contrasts were adjusted according to Scheffe (SAS Users' Guide, 2010).

To test whether more fractures or deviations where diagnosed when palpating or looking at radiographs, the assessment of the radiograph was deducted from the assessment by palpation (1 = presence of fracture of deviation, 0 = absence of fracture or deviation). A sign test was performed whether the result was unequal 0.

RESULTS

Inter-observer Repeatability

The inter-observer repeatabilities of the fractures and deviations of the whole keel (i.e., at least one fracture or deviation in at least one of the 3 locations) were improved by training and increased from R = 0.11 (fractures) and R = 0.18 (deviations) to R = 0.43 for both measures on the second day compared with the first day of the Training School (Table 1). The 95% confidence intervals for the fractures did not overlap at all and for deviations the estimate of 0.18 for the first day is not within the confidence interval of the estimate of the second day. There was evidence that the assessments of fractures and deviations of different parts of the keel bone had different repeatabilities. While the inter-observer repeatability of fractures at the cranial third of the keel bone including the 95% confidence interval was higher than 0 on the second day, the estimates of the inter-observer repeatabilities of fractures and deviations at the middle part of the keel bone were not significantly greater than 0 on either day and the estimates of the caudal parts of the keel bone

Table 1. Inter-observer repeatabilities before (day 1) and after training (day 2) based on the link scale. The mean repeatability estimates (R) and the 2.5% and 97.5% confidence intervals (CI) of the bootstrapping are given.

Variable	Time-point	R	CI
Fractures (N[1] = 394)[2]	Day 1	0.11	0.01, 0.24
Cranial (N = 131)[2]		0.11	0, 0.34
Middle (N = 124)[2]		0.11	0, 0.33
Caudal (N = 139)[2]		0.02	0, 0.12
Fractures (N = 420)[3]	Day 2	0.43	0.28, 0.58
Cranial (N = 141)[3]		0.33	0.07, 0.69
Middle (N = 138)[3]		0.04	0, 0.18
Caudal (N = 226)[3]		_[4]	_[4]
Deviations (N = 391)[2]	Day 1	0.18	0.06, 0.33
Cranial (N = 128)[2]		0.35	0.10, 0.61
Middle (N = 142)[2]		0.22	0, 0.73
Caudal (N = 121)[2]		0.02	0, 0.13
Deviations (N = 413)[3]	Day 2	0.43	0.27, 0.59
Cranial (N = 140)[3]		0.16	0, 0.69
Middle (N = 142)[3]		0.28	0.06, 0.51
Caudal (N = 121)[3]		_[4]	_[4]

[1]Number of observations
[2]Based on the assessments of 16 persons
[3]Based on the assessments of 15 persons
[4]Model failed to converge

Table 2. Frequencies of fractures and deviations as assessed from the radiographs during both days. The numbers refer to the assessments of all hens by each participant so each hen was assessed multiple times, but different hens were assessed on day 1 compared to day 2.

	Damage present	
Type of damage	Day 1	Day 2
Fracture, cranial	121 (55%)	90 (27.4%)
Fracture, middle	112 (52.6%)	81 (26.1%)
Fracture, caudal	246 (97.2%)	325 (95%)
Deviation, cranial	60 (36.8%)	61 (21.6%)
Deviation, middle	122 (63.9%)	222 (72.1%)
Deviation, caudal	20 (14.4%)	16 (6.4%)

(tip) were not significantly greater than 0 on day 1 and could not be estimated on day 2. Non-estimable repeatabilities could be due to low variances (https://cran.r-project.org/web/packages/rptR/vignettes/rptR.html). The confidence intervals of the inter-observer repeatabilities of deviations of the cranial and middle third of the keel bone were wide, often included 0 and overlapped on both days. The inter-observer repeatability of deviations at the caudal third was not significantly greater than 0 on day 1 and could not be estimated on day 2. The reason for the failure of estimation was the low variance in fractures and deviations at the caudal parts of the keel bones: Taking the assessments of radiographs as the best estimates of "true" incidences almost all caudal parts of the keel bone were fractured and hardly any had a deviation (Table 2).

Intra-observer Repeatability

The estimation of intra-observer repeatabilities of fractures and deviations was impaired because data on day 2 were only available from one person (Table 3).

Table 3. Intra-observer repeatabilities based on the link scale when keel bones were palpated twice by each observer. The mean repeatability estimates (R) and the 2.5 and 97.5% confidence intervals (CI) of the bootstrapping are given.

Variable	Time-point	R	CI
Fractures (N[1] = 340)[2]	Day 1	0.30	0.20, 0.40
Cranial (N = 116)[2]		0.13	0.04, 0.27
Middle (N = 101)[2]		0.02	0, 0.1
Caudal (N = 123)[2]		0.995	0.99, 0.997
Deviations (N = 306)[4]	Day 1	0.33	0.24, 0.42
Cranial (N = 100)[3]		0.07	0, 0.16
Middle (N = 117)[4]		0.16	0.001, 0.98
Caudal (N = 89)[4]		0.02	0, 0.09

[1]Number of observations
[2]Based on the assessments of 15 persons
[3]Based on the assessments of 13 persons
[4]Based on the assessments of 13 persons

Concentrating on day 1, the repeatability for fractures at the caudal part of the keel bone was very high whereas the repeatability for deviations at the caudal part of the keel bone was not significantly greater than 0. In general, repeatabilities of fractures were similar to the repeatabilities of deviations on day 1.

The intra-observer repeatability as measured with the binary variable "fracture/deviation outcome of the first palpation session vs. fracture/deviation outcome of the second palpation session" differed between assessors only for fractures of the cranial and the caudal part of the keel bone (cranial: Fisher's exact test: $P = 0.049$, N = 116, caudal: Fisher's exact test: $P = 0.003$, N = 123) and ranged from 33% (cranial fractures and deviations), 50% (caudal fractures), and 28.6% (caudal deviations) to 100% for fractures and deviations at all parts of the keel bone. In other words, the probability whether an assessor came to the same decision fracture/deviation yes/no differed during the first and the second palpation depended on the location (e.g., 50% of the assessors reported the same outcome for caudal fractures in both sessions). Assessors did not differ significantly concerning the intra-observer repeatability of deviations which ranged from 29% to 100% (all P values > 0.18).

Accuracy

The accuracy of palpations of fractures and deviations (i.e., accuracy of the palpation in session 2 compared with the assessment of the radiograph in session 3) were higher on day 2 than day 1 (Table 4). The accuracies of fractures at the middle and caudal part of the keel bones and of the deviations at all parts of the keel bone were especially high on day 2. The percentage of accurate assessments of fractures at different locations improved from 71% to 85% on day 1 to 78% to 96% on day 2 (Table 5a). Accuracy of palpation improved from day 1 to 2 for fractures at the middle and caudal parts of the keel bone, but no differences between the days were detected for fractures of the cranial part of the keel bone. Day ($F_{1, 894} = 8.74$, $P = 0.003$), location at the

Table 4. Accuracies of palpation, i.e., the agreement between the palpation and the "true" state as seen on the radiograph, based on the link scale. The mean repeatability estimates (R) and the 2.5% and 97.5% confidence intervals (CI) of the bootstrapping are given.

Variable	Time-point	R	CI
Fractures (N = 697)[1]	Day 1	0.42	0.33, 0.51
Cranial (N = 220)[1]		0.001	0, 0.01
Middle (N = 213)[1]		<0.0001	0, 0.0001
Caudal (N = 119)[1]		–[2]	–[2]
Fractures (N = 217)[3]	Day 2	0.61	0.52, 0.69
Cranial (N = 73)[3]		0.45	0.16, 0.96
Middle (N = 65)[3]		0.99	0.99, 0.998
Caudal (N = 79)[3]		0.98	0.96, 0.996
Deviations (N = 497)[1]	Day 1	0.41	0.32, 0.50
Cranial (N = 163)[1]		0.0002	0, 0.0001
Middle (N = 191)[1]		<0.0001	0, 0.0001
Caudal (N = 139)[1]		0.28	0, 0.996
Deviations (N = 199)[3]	Day 2	0.49	0.39, 0.59
Cranial (N = 66)[3]		0.99	0.99, 0.998
Middle (N = 74)[3]		0.99	0.98, 0.997
Caudal (N = 59)[3]		0.995	0.99, 0.998

[1]Based on the assessment of 15 people
[2]Model failed to converge.
[3]Based on the assessment of 10 people

Table 5. Here, accuracy is defined as the percentage of achieving the same assessment of keel bone damage from palpation and radiographs (i.e., session 2 vs. 3) in a binomial variable. The binary outcome variable was whether the assessment of fractures (a) and deviations (b) was identical between the palpation and at the assessment from the radiographs on day 1 and 2. The *P* values indicate if there was a significant training effect in the generalized linear model.

a) fractures Part	% day 1	% day 2	$F_{1,894}$	P
Cranial	77.4 (181)	78.1 (57)	0.04	0.84
Middle	71.1 (150)	90.8 (65)	7.88	0.005
Caudal	85.3 (215)	96.2 (76)	4.79	0.029
b) deviations				
Part	% day 1	% day 2	$F_{1,676}$	P
Cranial	77.1 (131)	89.4 (59)	5.63	0.02
Middle	82.8 (159)	85.1 (65)	1.05	0.31
Caudal	77.8 (105)	96.2 (76)	8.74	0.003

keel bone ($F_{2, 894} = 6.57$, $P = 0.002$) and the interaction between these 2 factors ($F_{2, 894} = 4.06$, $P = 0.018$) but not the assessor ($F_{14, 894} = 1.31$, $P = 0.19$) predicted the accuracy in the assessment of fractures. The difference in accuracy between the 2 days was different for the cranial part vs. the caudal part ($t_{894} = 3.61$, $P = 0.0003$) and the middle part vs. the caudal part ($t_{894} = -2.38$, $P = 0.018$) but not between the cranial vs. the middle part ($t_{894} = 1.24$, $P = 0.22$).

In contrast to fractures, accuracy of deviations depended on the assessors and a difference between days was present (Table 5b, day: $F_{1, 676} = 13.75$, $P = 0.0002$, assessor: $F_{14, 676} = 1.84$, $P = 0.03$). There was a trend that the difference between days depended on the location of the deviation (interaction day × location: $F_{2, 676} = 2.48$, $P = 0.09$). When this interaction was partitioned an effect of day was present for deviations of the cranial ($F_{1, 676} = 5.63$, $P = 0.02$) and caudal ($F_{1, 676} = 8.74$, $P = 0.003$) but not the middle part ($F_{1, 676} = 1.05$, $P = 0.31$) of the keel bone.

On average, assessors recorded more fractures and more deformations at all parts of the keel bones when palpating than when consulting the radiographs (all sign tests of fractures and deformations $0.0001 < P < 0.05$).

DISCUSSION

Confirming previous studies, training with radiographs enhanced the inter-observer repeatability when palpating keel bones of laying hens to diagnose fractures and deviations but this was different for different parts of the keel bone and different for fractures and deviations. The necessity for training (Wilkins et al., 2004; Petrik et al., 2013) arises because the advantages of the method of palpation to diagnose keel bone damage like speed and low invasiveness are counterbalanced by inaccuracies when compared with histology (Scholz et al., 2008) or dissection (Wilkins, et al., 2004; Stratmann et al., 2015).

This study revealed that repeatabilities between observers increased in the same magnitude for both keel bone fractures and deviations on the second day of training. Due to low sample size, the intra-observer repeatability of the second day cannot be interpreted. Assessors might have remembered some fractures or deviations from the preceding palpation which could have inflated the estimate of intra-observer repeatability. However, as each assessor recorded 6 results per bird (fracture, deformation for each of the 3 parts of the keel bone) for 10 hens before the same animal was palpated again this inflation is unlikely. Similarly to inter-observer repeatabilities, accuracy as defined as the agreement between the palpation and the assessment from the radiograph was higher on training day 2. In general, the repeatability values were quite low which might be due to the different degree on experience palpating keel bones among the observers. We cannot rule out that participants at the same station influenced each other even at times when they should not speak with each other. However, this would have inflated inter-observer repeatabilities similarly on days 1 and 2.

As in Buijs et al. (2018) inter-observer repeatability for fractures at the caudal part of the keel bone was very low on day 1. Possibly, the high frequency of fractures at the caudal part (above 95%) made the estimate of inter-observer repeatability difficult and probably was the reason why it could not be estimated on day 2. The high frequencies likely also inflated the intra-observer reliability estimates of fractures and deviations at the caudal part which had very high values. Therefore, these estimates of repeatabilities for palpation results concerning the caudal part of the keel bone might not be meaningful. In addition, most fractures at the caudal part were at the dorsal side of the keel bone which was inaccessible to palpation and could only be seen on the

radiographic image (Richards et al., 2011; Baur, pers. comm.). Furthermore, these fractures often miss callus, bony ectostosis, or suture material that could be felt during palpation (Casey-Trott et al., 2015). Especially for damage at the caudal part there seem to be limits to the accuracy of palpations that cannot be overcome by training. Using radiographic images (Rufener et al., 2018), CT's (Chargo et al., 2018), dissections, or, as the gold standard histology (Scholz et al., 2008) could enhance the quality of diagnosis.

Estimates of accuracies based on the percentages of identical assessments of palpation and radiographs, i.e., the binary outcome of "identical" or "not identical" scoring before and after radiograph assessment, were generally higher than the repeatabilities based on the intra-class correlation coefficient comparing palpations of session 1 and session 2. Regarding the applicability of palpation, one might argue that accuracy is a more relevant measure than inter-observer reliability. However, the accuracy of palpation depends heavily on the frequency of scores. As an example, the percentage of identical assessments of palpation and radiographs is automatically high when almost all caudal parts have multiple fractures that are more difficult to miss. Interestingly, assessors overestimated the number of fractures and deformations when they palpated compared to looking at radiographs. Recording a fracture or deformation after palpation but not recording it when consulting the radiographs happened more often than the opposite. It is important to note that this estimate of accuracy does not reveal the correct state but just how similar the same person assessed the same hen depending on the method. The aim of the training school was to improve the consistency between palpation and the assessment from the radiograph.

Multiple groups of assessors palpated different sets of birds which could have increased differences between assessors. However, care was taken to include the same range of damages in all sets so that this influence should be minimal. The main cause for the differences among assessors was probably their experience. At least one assessor regularly palpated hundreds of laying hens daily whereas others had never palpated a hen before. Unfortunately, we did not record the experience of each assessor so we could not take this into consideration in the analyses.

The necessity of training to obtain a scientifically valid palpation score is undisputed but commonly performed in different ways. Often, dissections of the keel bones are used to validate palpations during training (Buijs et al., 2018); Wilkins et al., 2004; Petrik et al., 2013) or CT scans with the construction of 3D models (Chargo et al., 2018) are used for training. In Buijs et al. (2018) one pre-trained assessor had compared his scoring to later dissected keel bones and the other pre-trained assessors had been instructed by experienced trainers. The present study describes the method of using radiographs for training purposes. However, it is important to note that even when looking at radiographs, the fracture or deviation status might not be clear although a radiologist experienced with keel bones was present and helped interpreting the radiographs. In several cases the radiologist was not certain if the image showed a fracture as only one latero-lateral image was provided, so some assessors might be more likely to diagnose a fracture or deviation from palpation than from the radiographic image. As a further complication of using radiographs to assess the true state, we cannot rule out that the previous palpation result influenced the evaluation of the radiographic image and inflated the measure of accuracy. Ultimately, indicators for impaired welfare due to fractures such as the pain experienced by the hen would be the relevant measure but this is even more difficult to assess.

In conclusion, training sessions with radiographic images improved inter-observer repeatability and accuracy and should be performed regularly before and during research projects involving palpation of keel bones. As a consequence, the palpation results of research groups might become more comparable.

ACKNOWLEDGEMENTS

The Training School was financed by the EU COST Action CA15224 Keel Bone Damage. We thank Lilian Smith for her valuable help before and during the event. Sarah Baur radiographed the hens and was available during the training sessions. Lastly, we thank all participants (B. Andersson, J. P. Christensen, T. Decroos, M. Guinebretiere, Z. Janjecic, S. Jansen, A. Jeremiasson, L. Jung, K. Kittelsen, U. Lenert, H. McCormack, M. Nasr, A. Pearson, N. Rokavec, B. Slavec, S. Werner) of the Training School who agreed to make their notes accessible for data analyses. Markus Schwab helped with the setup of the equipment.

REFERENCES

Bartlett, J. W., and C. Frost. 2008. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 31:466–475.

Bestman, M., and J.-P. Wagenaar. 2014. Health and welfare in dutch organic laying hens. Animals 4:374–390.

Buijs, S., J. L. T. Heerkens, B. Ampe, E. Delezie, T. B. Rodenburg, and F. A. M. Tuyttens. 2019. Assessing keel bone damage in laying hens by palpation: Effects of assessor experience on accuracy, inter-rater agreement and intra-rater consistency. Poult. Sci. 98:514–521.

Casey-Trott, T., J. L. T. Heerkens, M. Petrik, P. Regmi, L. Schrader, M. J. Toscano, and T. Widowski. 2015. Methods for assessment of keel bone damage in poultry. Poult. Sci. 94:2339–2350.

Casey-Trott, T. M., M. T. Guerin, V. Sandilands, S. Torrey, and T. M. Widowski. 2017. Rearing system affects prevalence of keel-bone damage in laying hens: A longitudinal study of four consecutive flocks. Poult. Sci. 96:2029–2039.

Chargo, N. J., C. I. Robison, S. L. Baker, M. J. Toscano, M. M. Makagon, and D. M. Karcher. 2018. Keel bone damage assessment: Consistency in enriched colony laying hens. Poult. Sci. 98:1017–1022.

Eusemann, B. K., U. Baulain, L. Schrader, C. Thöne-Reineke, A. Patt, and S. Petow. 2018. Radiographic examination of keel bone damage in living laying hens of different strains kept in two housing systems. PLoS One 13:e0194974.

Harlander-Matauschek, A., T. B. Rodenburg, V. Sandilands, B. W. Tobalske, and M. J. Toscano. 2015. Causes of keel bone damage and their solutions in laying hens. Worlds Poult. Sci. J. 71:461–472.

Heerkens, J. L. T., E. Delezie, B. Ampe, T. B. Rodenburg, and F. A. M. Tuyttens. 2016a. Ramps and hybrid effects on keel bone and foot pad disorders in modified aviaries for laying hens. Poult. Sci. 95:2479–2488.

Heerkens, J. L. T., E. Delezie, T. B. Rodenburg, I. Kempen, J. Zoons, B. Ampe, and F. A. M. Tuyttens. 2016b. Risk factors associated with keel bone and foot pad disorders in laying hens housed in aviary systems. Poult. Sci. 95:482–488.

Hester, P. Y., S. A. Enneking, B. K. Haley, H. W. Cheng, M. E. Einstein, and D. A. Rubin. 2013. The effect of perch availability during pullet rearing and egg laying on musculoskeletal health of caged White Leghorn hens. Poult. Sci. 92:1972–1980.

Käppeli, S., S. G. Gebhardt-Henrich, E. Fröhlich, A. Pfulg, H. Schäublin, and M. H. Stoffel. 2011a. Effects of housing, perches, genetics, and 25-hydroxycholecalciferol on keel bone deformities in laying hens. Poult. Sci. 90:1637–1644.

Käppeli, S., S. G. Gebhardt-Henrich, E. Fröhlich, A. Pfulg, and M. H. Stoffel. 2011b. Prevalence of keel bone deformities in Swiss laying hens. Br. Poult. Sci. 52:531–536.

Nakagawa, S., and H. Schielzeth. 2010. Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. Biol. Rev. Camb. Philos. Soc. 85:935–956.

Petrik, M. T., M. T. Guerin, and T. M. Widowski. 2013. Keel fracture assessment of laying hens by palpation: Inter-observer reliability and accuracy. Vet. Rec. 173:500–500.

Petrik, M. T., M. T. Guerin, and T. M. Widowski. 2015. On-farm comparison of keel fracture prevalence and other welfare indicators in conventional cage and floor-housed laying hens in Ontario, Canada. Poult. Sci. 94:579–585.

Regmi, P., N. Nelson, J. P. Steibel, K. E. Anderson, and D. M. Karcher. 2016. Comparisons of bone properties and keel deformities between strains and housing systems in end-of-lay hens. Poult. Sci. 95:2225–2234.

Riber, A. B., T. M. Casey-Trott, and M. S. Herskin. 2018. The influence of keel bone damage on welfare of laying hens. Front. Vet. Sci. 5:395–407.

Richards, G. J., M. Nasr, S. N. Brown, E. M. Gonzalez Szamocki, J. Murrell, F. Barr, and L. J. Wilkins. 2011. Radiography of laying hens. a useful tool for identifying keel bone fractures. Pages 49 in 5th International Conference on the Assessment of Animals. T. M. Widowski, P. Lawlis, and K. Sheppard, eds. Wageningen Academic Publishers.

Rodenburg, T. B., F. A. M. Tuyttens, K. de Reu, L. Herman, J. Zoons, and B. Sonck. 2008. Welfare assessment of laying hens in furnished cages and non-cage systems. An on-farm comparison. Anim. Welf. 17:363–373.

Rørvang, M. V., L. K. Hinrichsen, and A. B. Riber. 2019. Welfare of layers housed in small furnished cages on Danish commercial farms: the condition of keel bone, feet, plumage and skin. Br. Poult. Sci. 60:1–7.

Rufener, C., S. Baur, A. Stratmann, and M. J. Toscano. 2018. A reliable method to assess keel bone fractures in laying hens from radiographs using a tagged visual analogue scale. Front. Vet. Sci. 5:6–13.

SAS Institute Inc. 2010. SAS/STAT* User's Guide, Release 9.4 Edition. Cary, NC: SAS Institute Inc.

Schielzeth, H., and E. Bolund. 2010. Patterns of conspecific brood parasitism in zebra finches. Anim. Behav. 79:1329–1337.

Scholz, B., S. Rönchen, H. Harmann, M. Hewicker-Trautwein, and O. Distl. 2008. Keel bone condition in laying hens. a histological evaluation of macroscopically assessed keel bones. Berl. Munch. Tierarztl. Wochenschr. 121:89–94.

Scholz, B., S. Rönchen, H. Hamann, and O. Distl. 2009. Bone strength and keel bone status of two layer strains kept in small group housing systems with different perch configurations and group sizes. Berl. Munch. Tierarztl. Wochenschr. 122:249–256.

Stoffel, M. A., S. Nakagawa, H. Schielzeth, and S. Goslee. 2017. rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods Ecol. Evol. 8:1639–1644.

Stratmann, A., E. K. F. Fröhlich, S. G. Gebhardt-Henrich, A. Harlander-Matauschek, H. Würbel, and M. J. Toscano. 2015. Modification of aviary design reduces incidence of falls, collisions and keel bone damage in laying hens. Appl. Anim. Behav. Sci. 165:112–123.

Stratmann, A., E. K. F. Fröhlich, S. G. Gebhardt-Henrich, A. Harlander-Matauschek, H. Würbel, and M. J. Toscano. 2016. Genetic selection to increase bone strength affects prevalence of keel bone damage and egg parameters in commercially housed laying hens. Poult. Sci. 95:975–984.

Wilkins, L. J., S. N. Brown, P. H. Zimmerman, C. Leeb, and C. J. Nicol. 2004. Investigation of palpation as a method for determining the prevalence of keel and furculum damage in laying hens. Vet. Rec. 155:547–549.

Wilkins, L. J., J. L. McKinstry, N. C. Avery, T. G. Knowles, S. N. Brown, J. Tarlton, and C. J. Nicol. 2011. Influence of housing system and design on bone strength and keel bone fractures in laying hens. Vet. Rec. 169:414–414.