



## OPEN ACCESS

## EDITED BY

Rebecca E. Doyle,  
University of Edinburgh, United Kingdom

## REVIEWED BY

Jenny Yngvesson,  
Swedish University of Agricultural Sciences,  
Sweden  
Richard Blatchford,  
University of California, Davis, United States

## \*CORRESPONDENCE

Sarina Michaelis

✉ michaelis@uni-kassel.de

RECEIVED 04 May 2024

ACCEPTED 26 September 2024

PUBLISHED 01 November 2024

## CITATION

Michaelis S, Gieseke D and Knierim U (2024)  
Evaluating animal welfare on broiler and  
turkey farms using either normative  
values or descriptive benchmarks.  
*Front. Anim. Sci.* 5:1427733.  
doi: 10.3389/fanim.2024.1427733

## COPYRIGHT

© 2024 Michaelis, Gieseke and Knierim. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Evaluating animal welfare on broiler and turkey farms using either normative values or descriptive benchmarks

Sarina Michaelis\*, Daniel Gieseke and Ute Knierim

Farm Animal Behaviour and Husbandry Section, Faculty of Organic Agricultural Sciences, University of Kassel, Witzenhausen, Germany

For the welfare assessment on commercial broiler and turkey farms, not only the recording of animal-based indicators but also the evaluation of the resulting prevalence or rate is essential. Two evaluation methods were compared using data on welfare indicators collected over 1 year from 11 broiler and 11 turkey farms in Germany: the application of normative values from an evaluation framework and the calculation of a benchmark. The evaluation framework had recently been developed in a participatory process that provided an evaluation with target and alarm values. The target range was predominantly based on ethical considerations, while the alarm range was aligned with the current status quo from farm investigations. The 25th percentile and the 75th percentile of the benchmarking were similarly classified as target and alarm. When applying the evaluation framework across all indicators and flocks, 30.6% of broiler flocks were in the target range, while 41.4% were in the alarm range, mostly for indicators such as footpad dermatitis, weight uniformity, and mortality. For turkeys at week 5 or at the end of the fattening period, 51.6% and 32.9%, respectively, were in the target range and 12.3% and 14.4% were in the alarm range. Most alarm classifications were related to footpad dermatitis, low-weight uniformity, plumage damage, and skin injuries. The application of normative values led to a significantly worse average welfare rank over all indicators and flocks for broilers compared to the benchmark, while no difference was observed for turkeys. The farm selection process may have favored turkey farms with better management practices, resulting in a more rigorous benchmark than in broilers. In addition, the farm data used to set the normative values had indicated a poorer status quo in turkeys for certain indicators, resulting in less stringent limits for the alarm range. This highlights the challenges associated with both evaluation methods: normative values are affected by the process and criteria used to set them, while benchmarks are affected by the reference population, which calls for large databases with regular updates. Also, for normative values, developments in the sector and the latest scientific evidence should be used for recurrent validation.

## KEYWORDS

self-assessment, welfare protocol, animal-based indicators, benchmark, normative values, prevalences, welfare assessment, welfare evaluation

## 1 Introduction

Animal welfare is composed of different aspects that determine an animal's quality of life. Thus, it is a multifaceted or multidimensional state that can range from very poor to very good. The multitude of relevant aspects can be categorized in different ways, essentially covering the physical and mental state of the individual animal (TierSchG, 2006; World Organisation for Animal Health, 2024). For the assessment of animal welfare, therefore, no single indicator can be used alone, but a larger number of measures should be applied, depending on the question being asked (Fraser, 1995; Knierim et al., 2001). These measures can be animal-based on the one hand or resource- or management-based on the other. While both types have advantages and disadvantages, animal-based indicators provide direct and therefore more valid information about the animal's state (Knierim and Winckler, 2009). Welfare assessment on commercial farms includes, as a first step, the recording of data by qualified assessors that indicate the physical and mental state of the herd or flock as a whole, mainly by quantifying the proportion of animals with specific welfare problems while covering different dimensions of animal welfare. For broilers (*Gallus gallus dom.*) and turkeys (*Meleagris gallopavo dom.*), a number of protocols with predominantly animal-based indicators are available, such as the Welfare Quality<sup>®</sup> protocol for broilers (Welfare Quality<sup>®</sup> Consortium, 2009), the AWIN protocol for turkeys (Estevez et al., 2015), or the KTBL guidelines for farmers' self-assessment of broiler and turkey welfare (Knierim et al., 2020a). Coordinated by the KTBL (German Association for Technology and Structures in Agriculture), the latter have been developed through a participatory process involving experts from farming practice, extension, competent authorities, animal welfare organizations, and research. Animal-based welfare indicators have been selected that validly address the most important animal welfare issues known from practice and which are suitable for reliable use by farmers after a short training (Zapf et al., 2015). Important welfare issues are for example the occurrence of footpad dermatitis in broilers (e.g., Allain et al., 2013) and turkeys (e.g., Freihold et al., 2019; Leishman et al., 2021), of skin injuries and feather damage in turkeys (e.g., Allain et al., 2013), and of lameness in broilers (e.g., Granquist et al., 2019) and in turkeys (e.g., Ferrante et al., 2019). The welfare self-assessment is designed to help farmers improve their animal management by identifying welfare problems and monitoring the effectiveness of preventive measures. In order to draw conclusions from the application of the protocols, in a second step, it is necessary to evaluate for each indicator the degree to which the data indicate a desirable welfare state or a welfare problem. When we use the term welfare evaluation in the following, we refer to this second step. Two main evaluation approaches are possible: comparison with normative values that define the levels of acceptable or unacceptable welfare, and benchmarking, which compares individual farm outcomes with those of a sample of other farms, often with the upper or lower quartiles of the reference population.

One example of a normative approach is the evaluation framework elaborated for the Welfare Quality<sup>®</sup> protocols. It is based on expert opinion and translated into algorithms to calculate single and aggregated indicator scores allocated to four welfare

categories from "not classified" to "excellent" (Botreau et al., 2009). The expert opinion was based on normative considerations, namely, on theoretical goals, but in relation to "what can realistically be achieved in practice" (Botreau et al., 2009). This combination is rather common for the setting of animal welfare targets in farm animals; nevertheless, only the term normative will be used in the following. Normative values can also be found in European or national legislation. For instance, the EU Directive, 2007/43/EC for the protection of chickens kept for meat production (EU Directive, 2007) stipulates a maximum acceptable cumulative daily mortality rate (1% plus 0.06% for every day of life) in at least seven consecutive flocks from a stable in case that an increased stocking density is used. The German implementing provisions relating to this directive further defined thresholds of 0.5% for birds dead on arrival and 40% for superficial or 20% for severe footpad dermatitis recorded at the slaughterhouse (Implementing Provisions TierSchNutztV, 2014).

Benchmark approaches are often established in vertical integrations or in food quality schemes where farm data can be centrally processed by a company such as Heidemark GmbH (Heidemark, 2023) or Wiesenhof Geflügel-Kontor GmbH in Germany. Another widely known example for the use of benchmarking is the monitoring of farm-level antimicrobial use applied in many countries (Sanders et al., 2020). An essential component of benchmarking is the implementation of a digital infrastructure, facilitating the submission, storage, and processing of farm data.

There are advantages and disadvantages to both approaches. A normative framework allows each individual farm's outcomes to be assessed on a level playing field, whereas benchmarking is influenced by the size and selection of the reference population. For instance, the reference values related to the AWIN turkey welfare assessment protocol are derived from 44 flocks from 26 farms (Estevez et al., 2015) which may raise questions on generalizability. On the other hand, the setting of normative values can meet with resistance from the sector, especially if they have not been involved in the process and see the values as unattainable. Benchmarks are more status-quo-oriented. They inform farmers about their performance compared to their peers. Results in the poorer range may motivate farmers, for example, to improve weaknesses (Main et al., 2003; von Keyserlingk et al., 2012; Atkinson et al., 2017; Sumner et al., 2020), which can lead to better welfare (Pandolfi et al., 2017). At the same time, for severe welfare problems, even the best farm outcome can be ethically unacceptable, or at the other extreme, the worst quarter may still be in the acceptable range if the welfare level is generally high.

The German Animal Welfare Act (TierSchG, 2006) requires farmers to perform self-assessments of animal welfare on their farms using animal-based indicators. The KTBL guidelines (Knierim et al., 2020a) are one proposal on how to implement the self-assessments on farms. However, not only a measurement of indicators is required, but also an evaluation of the outcomes. Therefore, a multistage process was used for the development of normative values to help farmers determine whether the welfare of their birds in relation to each indicator is satisfactory or whether action needs to be taken.

We present the resulting normative values, which we apply to welfare self-assessments on commercial broiler and turkey farms and compare the evaluation outcomes with those from a benchmarking approach. We hypothesize that applying the evaluation framework to the farms' rates and prevalences would result in a more negative evaluation than the results achieved through a benchmarking system.

## 2 Materials and methods

### 2.1 Development of the evaluation framework for the self-assessment

The multistage process of developing the evaluation framework began in 2018 and comprised two rounds of a Delphi survey and three expert discussions based on the results of the Delphi survey and status quo information from German on-farm studies that had assessed broiler or turkey welfare using methods similar to the self-assessment protocols. The Delphi survey was coordinated by KTBL for poultry (also including laying hens) and additionally for cattle and pigs (Zapf et al., 2023; Schultheiß et al., 2023). All known German poultry experts from expert directories and from snowball sampling were invited to participate in the Delphi survey and were additionally asked to forward the survey to suitable colleagues. This included farmers, advisors and veterinarians, scientists, and representatives from competent authorities and farmers' and animal welfare organizations. Participants were asked to propose "target values" and "alarm values" representing a "traffic light" evaluation (for more information, see Schultheiß et al., 2023). The target value was limiting the target (green) range. It was explained that it should be achievable under farm conditions, e.g., that a certain tolerance above 0% of animals with welfare problems should be considered. The range between the target value and the alarm value was regarded as an early-warning (yellow) range. The alarm value was the threshold to a problematic (red) area where immediate action would need to be taken. The participants provided values for each indicator and, after receiving a summary of all answers, had the option to amend their proposal. The response rate from the 161 contacted poultry experts was 19%, with one-third being scientists. No farmer responded.

In parallel, information about the current status quo in relation to the different welfare indicators was sought from on-farm studies of broiler and turkey welfare. Results from five German studies on commercial farms were available that had applied comparable assessment methods and could provide raw data on rates or prevalences (Westermaier, 2015; Rösler, 2016; Hübel, 2019; Olschewsky, 2019; Toppel, 2020). Percentiles (25th, 50th, and 75th) were calculated. They were weighted according to the number of flocks if more than one publication provided data.

On this basis, three rounds of expert discussions were held over 3 years, with 24 participants in the first round and 20 in the last. The poultry experts were farmers, advisors and veterinarians, scientists, and representatives from competent authorities and farmers' and animal welfare organizations. Given the limited number of participants, they had been selected from expert directories in

order to represent all relevant stakeholders in the poultry sector and also to ensure a diversity of viewpoints. It was possible to participate in both the Delphi survey and the expert discussions. In the discussions, however, farmers and farmer representatives also participated. Proposals for target and alarm values for each indicator were discussed, taking into account the results of the Delphi survey and the percentiles from the five on-farm studies (Westermaier, 2015; Rösler, 2016; Hübel, 2019; Olschewsky, 2019; Toppel, 2020). In each discussion round, a consensus was sought, but participants could also submit comments on the values afterward.

### 2.2 Farms and farm visits

On-farm self-assessments were carried out by 11 commercial broiler and 11 commercial turkey farms in Germany. They were recruited through advertisements in the agricultural press and direct contact with farmers ( $n = 11$ ) as well as through recommendations from farmers' associations or industry ( $n = 11$ ). The sample of farms was intended to reflect the range of typical German farms. The farm sizes of the participating farms varied between 1,680 and 97,000 broilers and 2,000 and 25,200 turkeys. The majority were conventional farms using Ross 308 and BUT 6 genetics. Three organic farms raised ISA JA 757 broilers (farm size 1,680 to 4,800), and two organic turkey farms reared Kelly bronze turkeys (farm size 2,000 to 13,500). One turkey farm had both conventional and organic flocks. All broiler farms had mixed flocks of female and male birds. Three turkey farms kept mixed flocks until about the fifth week of life (one organic and two conventional farms). Six farms had male (one organic and five conventional farms) and two had female flocks (conventional farms).

The self-assessments were conducted in accordance with the first edition of the KTBL guidelines (Knierim et al., 2016) over a period of 1 year. The applied indicators and their definitions are displayed in Tables 1, 2. In 2020, the KTBL guidelines were updated to a second edition, which is available online (Knierim et al., 2020a).

The test year started at individual time points between December 2018 and August 2019 after the farmers had completed an online or in-person training with a reliability test (for details on training and reliability testing, see Michaelis et al., 2022). One turkey farm withdrew during the study due to time constraints of the livestock owner. Each farm was visited by the first or second author during the first and second halves of the test year to record all the indicators to be assessed on individual birds (Table 1). A sample of 50 birds was assessed in parallel with the farmer, and additionally, these birds were weighed for the indicator weight uniformity (Table 2). In compliance with the recommended assessment schedule by the KTBL guidelines, on each farm, the samples of broilers were once assessed in about the second and once in the last week of life and turkeys once in the fifth week and once in the last month of life. The first farm visit was not necessarily an assessment of a young flock and the two visits were always on different flocks. Indicators assessed included footpad dermatitis, lameness, and weight uniformity in broilers; lameness was only assessed at the end of the fattening period (Table 1). For turkeys, the

**TABLE 1** Definitions of indicators of the KTBL-guideline (Knierim et al., 2016) measured on individual animals.

Indicator	Species	Definition
Extent and quality of beak trimming	Turkey	0: Beak is not trimmed 1: Moderately trimmed upper beak (no opening visible from the front, beak can be closed) 2: Severely trimmed upper beak (opening visible from the front, beak cannot be closed)
Plumage condition	Turkey	0: No featherless areas 1: Small featherless areas (one or more featherless areas <5 cm longest diameter) 2: Large featherless areas (at least one area ≥5 cm longest diameter)
Skin injuries	Turkey	0: No pecking or scratching injuries 1: Slight pecking or scratching injuries <1 cm length or diameter on snood or wattle or <2 cm length or diameter on head, back, wings 2: Moderate to severe pecking or scratching injuries ≥1 cm length or diameter on snood or wattle or ≥2 cm length or diameter on head, back, wings
Footpad dermatitis	Broiler	2nd week of life: 0: No to very slight alterations (slight coloration of scales or in between, no deep lesions) 1: Hyperkeratosis or deep lesions with damage of the upper skin layer  Last week of life: 0: Intact skin (at max. slight hyperkeratosis, no coloration or scars) 1: Superficial lesions (or erosions), hyperkeratosis, and coloration 2: Severe, deep lesions, ulcer and scars mostly circular at the metatarsal footpad or metatarsal and digital pads equally affected
	Turkey	0: Intact footpad or not more than punctual necrosis on the metatarsal pad with no swelling and toes not affected 1: Extensive necrosis on up to 1/2 of the metatarsal pad and swelling or scabs or toes slightly affected 2: Necrosis on more than 1/2 of the metatarsal pad or strong swelling or scabs, deep injuries, toes affected
Lameness	Broiler, turkey	0: Normal gait or slightly affected (no or small defects, e.g., overlong strides or unsteady gait) 1: Clear gait defect, e.g., severe limping, jerking on one leg, abducted leg during walking, sitting down repeatedly after stopping (within 2 s), severely impaired maneuverability, bird is incapable to walk durably

**TABLE 2** Definitions of indicators of the KTBL-guideline (Knierim et al., 2016) based on flock data.

Indicator	Species	Definition
Weight uniformity	Broiler, turkey	Percentage of birds within ±10% of the mean weight of weighted birds
Mortality rate at first week of life	Broiler	Cumulative rate of dead or culled birds in the first week of life, based on the daily number of animals
	Turkey	Percentage of dead or culled birds in the first week of life, based on the number of animals on day 1
Weekly mortality rate	Broiler	Cumulative rate of dead or culled birds in a week, based on the daily number of animals
	Turkey	Rate of dead or culled birds in a week, based on the number of animals on day 1
Total mortality rate	Broiler	Cumulative rate of dead or culled birds based on the daily number of animals
	Turkey	Rate of dead or culled birds over the rearing or fattening period based on the number of animals on day 1
Animals dead on arrival at the slaughterhouse	Broiler, turkey	Percentage of birds dead on arrival at the slaughterhouse after transport and catching
Condemnation of birds at the slaughterhouse	Broiler, turkey	Percentage of condemnations per flock

indicators assessed were the extent and quality of beak trimming, plumage condition, skin injuries, footpad dermatitis, lameness, and weight uniformity (Table 2).

For farms with multiple barns, the farmer selected one barn to participate in the study. On three broiler farms, two flocks were kept concurrently in the selected barn. For these, the sample of 50 birds for the assessment of the indicators was taken proportionally from each flock. This also applied to the turkey flocks in the fattening period, which were kept as mixed flocks during rearing. However, the mortality rates and slaughterhouse indicators were calculated separately for concurrently kept flocks and the subsequently separated mixed turkey flocks. For the turkey farm with both conventional and organic flocks, the selected barn changed from a conventional barn to an organic barn between the two farm visits.

Prior to the farm visits, the first and the second authors had been trained on-farm by three different experienced scientists, working with animal-based indicators in broilers and turkeys (of whom one was an author of the KTBL guidelines). Both authors

were experienced in the application of the assessment protocol due to several interrater reliability tests with other scientists, resulting in substantial to almost perfect agreement for both regarding all indicators assessed on individual birds [broilers: prevalence-adjusted and bias-adjusted kappa (PABAK) = 0.88 to 1.00, *n* = 65; turkeys: 0.65 to 1.00, *n* = 60].

### 2.3 Data processing, evaluation, and analysis

The flock data provided by the farmers during the test year were based on the mortality records of the farmers and the slaughterhouse records on condemnations and birds dead on arrival. The farmers either submitted unprocessed records or extracted data in Excel sheets and sent them by post or email. Both the number of flocks per farm during the test year and the completeness of the submitted data varied, resulting in data from altogether 40 to 82 flocks per indicator in broilers and from 7 to 31 flocks in turkeys. From the mortality records, mortality in the first week of life, weekly mortality, and total mortality were calculated. Furthermore, rates for birds dead on arrival at the slaughterhouse and condemnation of birds were calculated (broilers: *n* = 278, Table 3; turkeys: *n* = 157, Table 4). Prevalences were derived from the authors' assessments during the two farm visits (broilers: *n* = 64, based on three indicators applied on 11 farms at two visits, Table 3; turkeys: *n* = 187, based on six indicators applied during 10 to 11 farms at two visits, Table 4). Due to

TABLE 3 Prevalences and rates (in %) of each indicator reached by the broiler farms and the number of farms and flocks per indicator.

Indicator	Prevalences or rates of the participating farms		
	Mean (minimum – maximum)	<i>n</i>	
	All (score 1 + 2)	Severe (score 2)	Farms/ flocks
Footpad dermatitis (2nd week of life)	12.7 (0.0–52.0)	n.a.	11/11
Footpad dermatitis (last week of life)	47.1 (0.0–100.0)	11.8 (0.0–70.0)	11/11
Lameness (last week of life)	4.7 (0.0–10.0)	n.a.	11/11
Weight uniformity (2nd week of life)	58.7 (35.0–74.0)	n.a.	11/11
Weight uniformity (last week of life)	50.4 (32.0–66.0)	n.a.	9/9
Mortality first week of life	0.93 (0.15–3.06)	n.a.	7/58
Weekly mortality	0.42 (0.17–1.22)	n.a.	7/58
Total mortality rate	3.16 (0.16–14.99)	n.a.	10/82
Dead on arrival at the slaughterhouse	0.07 (0.0–0.32)	n.a.	6/40
Condemnation of birds	0.99 (0.0–4.62)	n.a.	6/40

n.a., not applicable.

TABLE 4 Prevalences and rates (in %) of each indicator reached by the turkey farms and the number of farms and flocks per indicator.

Indicator	Prevalences or rates of the participating farms		
	Mean (minimum–maximum)	<i>n</i>	
	All (score 1 + 2)	Severe (score 2)	Farms/ flocks
Extent and quality of beak trimming	2.9 (0.0–18.0)	n.a.	11/21
Plumage condition (r)	0.0 (0.0–0.0)	0.0 (0.0–0.0)	10/10
Plumage condition (f)	18.9 (0.0–62.0)	3.1 (0.0–14.0)	11/11
Skin injuries (r)	24.4 (8.0–48.0)	2.0 (0.0–8.0)	10/10
Skin injuries (f)	27.3 (14.0–58.0)	4.7 (0.0–8.0)	11/11
Footpad dermatitis (r)	39.0 (4.0–90.0)	0.0 (0.0–0.0)	10/10
Footpad dermatitis (f)	72.4 (22.0–100.0)	4.4 (0.0–24.0)	11/11
Lameness (r)	0.0 (0.0–0.0)	n.a.	10/10
Lameness (f)	0.2 (0.0–2.0)	n.a.	11/11
Weight uniformity (r)	69.3 (50.0–81.0)	n.a.	9/9
Weight uniformity (f)	77.7 (60.0–90.0)	n.a.	10/10
Mortality first week of life	0.86 (0.31–1.37)	n.a.	4/11
Weekly mortality (r) Female and male flocks	♀: 0.43 (0.23–0.62) ♂: 0.55 (0.46–0.68)	n.a.	8/11
Total mortality rate (r) Female and male flocks	♀: 2.13 (1.13–3.08) ♂: 2.85 (2.31–3.41)	n.a.	8/11
Weekly mortality (f) Female flocks	0.20 (0.10–0.30)	n.a.	3/7
Weekly mortality (f) Male flocks	0.33 (0.10–0.56)	n.a.	5/22
Total mortality rate (f) Female flocks	2.05 (0.10–3.0)	n.a.	3/7
Total mortality rate (f) Male flocks	5.68 (1.32–8.73)	n.a.	7/26
Dead on arrival at the slaughterhouse	0.11 (0.0–0.36)	n.a.	8/31
Condemnation of birds	1.12 (0.0–3.47)	n.a.	8/31

r, rearing; f, fattening; n.a., not applicable.

technical issues, weight measurements were missing from farm visits of two broiler and two turkey flocks.

These prevalences and rates were further evaluated using the target and alarm values of the evaluation framework, considering the severity of alterations (total score 1 + 2; severe score 2), sex (for weekly and total mortality in turkeys), and the time point of assessment (Tables 5, 6). Additionally, the 25th and 75th percentiles were calculated for the benchmark.

**TABLE 5** Normative values of the evaluation framework for broilers (Knierim et al., 2020b), medians from the Delphi survey and percentages of the flocks categorized into target, early-warning or alarm range.

Indicator	Normative values of the evaluation framework (versus median values from the Delphi survey)				Percentages of flocks in the target/early-warning/ alarm range	
	All (score 1 + 2)		Severe (score 2)		All (1 + 2)	Severe (2)
	Target	Alarm	Target	Alarm		
Footpad dermatitis (2nd week of life)	0.0 (0.5)	6.0 (5.0)	n.a.	n.a.	46/18/36	n.a.
Footpad dermatitis (last week of life)	8.0 (4.3)	24.0 (17.5)	2.0 (1.0)	6.0 (5.5)	18/18/64	46/18/36
Lameness (last week of life)	0.0 (1.0)	4.0 (6.8)	n.a.	n.a.	36/18/46	n.a.
Weight uniformity (2nd week of life)	85.0 (85.0)	60.0 (70.0)	n.a.	n.a.	0/54/46	n.a.
Weight uniformity (last week of life)	85.0 (85.0)	50.0 (70.0)	n.a.	n.a.	0/44/56	n.a.
Mortality first week of life	0.5 (n.a.)	0.9 (n.a.)	n.a.	n.a.	22/28/50	n.a.
Weekly mortality	0.21 (n.a.)	0.35 (n.a.)	n.a.	n.a.	9/27/64	n.a.
Total mortality rate	0.5 + 0.03 × dol (1.5)	0.9 + 0.05 × dol (4.5)	n.a.	n.a.	35/24/41	n.a.
Dead on arrival at the slaughterhouse	0.05 (0.3)	0.3 (0.5)	n.a.	n.a.	45/53/2	n.a.
Condemnation of birds	0.5 (0.5)	1.5 (1.5)	n.a.	n.a.	60/15/25	n.a.

dol, day of life; n.a., not applicable.

**TABLE 6** Normative values of the evaluation framework for turkeys (Knierim et al., 2020c), medians from the Delphi survey, and percentages of the flocks categorized into target, early-warning, or alarm range.

Indicator	Normative values of the evaluation framework (versus median values from the Delphi survey)				Percentages of flocks in the target/ early-warning/alarm range	
	All (score 1 + 2)		Severe (score 2)		All (1 + 2)	Severe (2)
	Target	Alarm	Target	Alarm		
Extent and quality of beak trimming	2.0 (0.5)	6.0 (2.0)	n.a.	n.a.	r: 90/10/0 f: 82/0/18	n.a.
Plumage condition (r)	4.0 (2.5)	8.0 (6.5)	2.0 (0.3)	6.0 (2.0)	100/0/0	100/0/0
Plumage condition (f)	6.0 (6.0)	30.0 (19.0)	2.0 (1.0)	10.0 (5.0)	37/36/27	82/9/9
Skin injuries (r)	4.0 (0.8)	20.0 (4.0)	2.0 (0.3)	6.0 (1.0)	0/40/60	70/10/20
Skin injuries (f)	6.0 (5.0)	30.0 (10.0)	2.0 (0.8)	6.0 (2.8)	0/73/27	27/27/46
Footpad dermatitis (r)	2.0 (1.0)	40.0 (5.0)	0.0 (0.0)	4.0 (1.0)	0/70/30	100/0/0
Footpad dermatitis (f)	30.0 (4.0)	60.0 (10.0)	2.0 (1.0)	6.0 (2.0)	9/27/64	82/0/18
Lameness (r)	0.0 (0.0)	4.0 (2.0)	n.a.	n.a.	100/0/0	n.a.
Lameness (f)	0.0 (2.0)	4.0 (5.0)	n.a.	n.a.	91/9/0	n.a.
Weight uniformity (r)	90.0 (90.0)	70.0 (77.5)	n.a.	n.a.	0/56/44	n.a.
Weight uniformity (f)	90.0 (92.5)	70.0 (80.0)	n.a.	n.a.	10/60/30	n.a.
Mortality first week of life	0.6 (n.a.)	1.7 (n.a.)	n.a.	n.a.	45/55/0	n.a.
Weekly mortality (r) Female and male flocks	0.2 (n.a.)	0.7 (n.a.)	n.a.	n.a.	♀: 0/100/0 ♂: 0/100/0	n.a.

(Continued)

TABLE 6 Continued

Indicator	Normative values of the evaluation framework (versus median values from the Delphi survey)				Percentages of flocks in the target/early-warning/alarm range	
	All (score 1 + 2)		Severe (score 2)		All (1 + 2)	Severe (2)
	Target	Alarm	Target	Alarm		
Total mortality rate (r) Female and male flocks	1.4 (n.a.)	4.5 (n.a.)	n.a.	n.a.	♀: 29/71/0 ♂: 0/100/0	n.a.
Weekly mortality (f) Female flocks	0.11 (n.a.)	0.35 (n.a.)	n.a.	n.a.	14/86/0	n.a.
Weekly mortality (f) Male flocks	0.2 (n.a.)	0.5 (n.a.)	n.a.	n.a.	18/73/9	n.a.
Total mortality rate (f) Female flocks	0.11 × wol (n.a.)	0.35 × wol (n.a.)	n.a.	n.a.	29/71/0	n.a.
Total mortality rate (f) Male flocks	0.2 × wol (n.a.)	0.5 × wol (n.a.)	n.a.	n.a.	27/69/4	n.a.
Dead on arrival at the slaughterhouse	0.05 (0.1)	0.5 (0.5)	n.a.	n.a.	19/81/0	n.a.
Condemnation of birds	0.5 (0.5)	2.0 (2.0)	n.a.	n.a.	22/68/10	n.a.

r, rearing; f, fattening; wol, week of life; n.a., not applicable.

They were ranked into three ranges: less than or equal to the target value or 25th percentile (target range; coded as 1), between the target and alarm value or 25th and 75th percentiles (early-warning range; coded as 2), and above the alarm value or 75th percentile (alarm range; coded as 3). Note that for weight uniformity, the ranges are reversed. For the comparison of the normative and the benchmark evaluation, the resulting ranks (1, 2, 3) of the recorded indicators per farm were used. The indicators were subdivided by severity for those assessed on a sample of birds and by sex for weekly and total mortality in turkeys. Thus, up to 10 ranked measurements were possible for a broiler farm, 14 for a turkey farm raising only one sex, and 16 for a mixed-sex turkey farm. To prevent pseudoreplication, a mean rank was calculated when multiple measurements of the same indicator were available from successive flocks assessed at a farm during the test year. The number of flocks within a mean ranged from 2 to 14 over all indicators. Data were unavailable for 19 measurements in broilers and 39 in turkeys (including missing mortality measurements of the other sex in farms keeping one-sex flocks). As a result, 91 ranked measurements for broilers and 137 for turkeys were paired with the equivalent benchmarks for comparison. A two-tailed Wilcoxon signed-rank test (due to non-normally distributed ranks, Shapiro–Wilk test) was used to compare the two evaluation methods, with the pairs being the statistical unit. All calculations were conducted with the statistical software R (version 4.2.2).

## 3 Results

### 3.1 The evaluation framework

The final agreement among the participating stakeholders on the evaluation framework with normative values was reached under the condition that the target values should be guided by the results of the

Delphi survey, while the alarm values should more closely reflect what appears to be achievable in common practice, namely, guided by the 50th percentiles of the compiled raw data of the farm studies considered. It was noted that for indicators assessed in a sample of 50 birds (according to the recommendations of Knierim et al., 2016, 2020a), prevalences can only occur in 2% increments. Therefore, it was also agreed that a minimum difference of 4% between the target and alarm value (rounded in 2% increments if necessary) should be maintained to provide an early-warning range. Therefore, the framework values partly differed with respect to the values proposed by the Delphi survey, as indicated in Tables 5, 6. The evaluation framework is published in German (Knierim et al., 2020b, c).

### 3.2 The results of the welfare assessments and welfare evaluations using the evaluation framework

The recorded prevalences and rates relating to all indicators in the self-assessment protocol are displayed in Tables 3, 4.

According to the evaluation framework for broilers (Table 5), 30.6% of all flocks over all indicators were in the target range and 41.4% were in the alarm range. Considering only the indicators assessed on a sample of birds (footpad dermatitis, lameness, and weight uniformity), 25.0% of the flocks were in the target range, while 48.4% were in the alarm range. For total mortality, 34.9% of the flocks were in the target range and 41.0% in the alarm range. For the slaughterhouse indicators, 52.5% of the flocks were in the target range and 13.6% exceeded the alarm value.

According to the evaluation framework for turkeys (Table 6), approximately half of the flocks over all indicators (51.6%) were in the target range during the rearing period, and 12.3% of the flocks were in the alarm area. At the end of the fattening period, 32.9% of the flocks were in the target range and 14.4% in the alarm range. For

indicators assessed on a sample of birds (extent and quality of beak trimming, plumage condition, skin injuries, footpad dermatitis, lameness, and weight uniformity), during the rearing period, 62.9% of the flocks were in the target range, decreasing to 46.9% at the end of the fattening period. For mortality during the rearing period, 100% of the male flocks were in the early-warning range, falling to 70.8% during the fattening period, with 22.9% of the flocks in the target range and 6.3% in the alarm range. In female flocks, 85.7% were in the early-warning range during rearing and none in the alarm range, falling to 78.6% in the early-warning range during fattening with again none in the alarm range. For the two slaughterhouse indicators dead on arrival and condemnation rate, 21.0% of the flocks were in the target range, 74.2% in the early warning, and 4.8% in the alarm range.

### 3.3 Comparison of the evaluations using normative values or benchmarks

The comparison of evaluation results using normative values versus the benchmark approach showed significant differences for the broiler farms ( $n_{\text{pairs}} = 91$ ;  $p = 0.002$ ,  $W = 825.5$ ). The median rank over all farms and indicators was 2.0 for both approaches; however, the mean was 1.9 for the benchmarking and 2.1 for the evaluation framework. For the turkey farms, the evaluations led to similar results between both approaches ( $n_{\text{pairs}} = 148$ ;  $p = 0.699$ ,  $W = 892.0$ ), with a median and mean of 2.0 for the normative evaluation and the benchmarking. Thus, on average, the broiler farms were rated slightly worse in the normative evaluation, while this difference was not present for turkey farms.

## 4 Discussion

The process to develop a normative evaluation framework for the farmers' self-assessment of animal welfare was initially met with resistance by the poultry sector which was reflected by the non-participation of farmers in the Delphi survey. Farmers were concerned that the normative values would not only serve their own evaluations but also be used by authorities to identify infringements of animal welfare legislation. In addition, the survey was very extensive, with many different indicators and scores to consider. This may have further contributed to the low response rate of 19%. Considering these non-responders, it can be expected that the Delphi results were biased toward higher ambitions for poultry welfare. However, finally, in the expert discussion, the industry representatives did take part and influenced the negotiation process. An agreement was reached not only considering ethical considerations but also the current status quo in practice using mostly the 50th percentile of recent prevalences from German farms as guidance. Botreau et al. (2009) reported a similar strategy for the Welfare Quality® evaluation based on expert opinion. However, for welfare problems that are still widespread in practice with many birds affected, the adjustments to the 50th percentiles had led to large increases of alarm values compared to the originally proposed values from the

Delphi survey, such as for plumage condition, skin injuries, and footpad dermatitis in turkeys where the alarm value was increased by a factor of up to 8. For footpad dermatitis in turkeys, even the target value was increased due to the high prevalences in the compiled study data.

Nevertheless, when applying the evaluation framework to the assessment data of the participating farms, the results across all indicators and flocks of both species showed that approximately two-thirds were outside the target range. For the turkeys during the rearing period, the evaluation results were better with approximately half of the results outside the target range. However, except for footpad dermatitis, only a minority of turkey flocks were in the alarm range. In contrast, broiler flocks were more often evaluated to be in the alarm range than in the early-warning range. One reason for this difference between broiler and turkey flocks may be a bias in the study farms toward turkey farms with better management and, consequently, a better bird welfare. Another reason was the lower alarm values set for broilers compared to turkeys, such as 24% for total footpad dermatitis in broilers versus 60% in turkeys. This difference resulted from higher reported prevalences in turkeys than broilers found in the study data on the welfare state on commercial farms used during the agreement process (Westermaier, 2015; Rösler, 2016; Hübel, 2019; Olschewsky, 2019; Toppel, 2020).

The question arises whether the prevalences and rates measured across the varying number of flocks in our study, along with the resulting low percentages of welfare outcomes within the target range, accurately reflect the typical welfare levels in broiler and turkey farming. While currently, to our knowledge, no similar evaluations are available in the international literature, we can compare the prevalences and rates of the participating farms with literature values obtained by similar methods under comparable husbandry conditions or with legal thresholds. As our recorded prevalences are based on conventional and a few organic farms with fast- and slow-growing genetics, we consider studies on conventional and organic, fast- and slow-growing birds to determine the ranges found in practice. Furthermore, in terms of the acceptability of prevalences and rates with respect to bird welfare, the conditions under which impairments occur should play little or no role.

Regarding footpad dermatitis in turkeys, the recorded mean prevalence of 72% at the end of the fattening period (scores 1 and 2) was only slightly higher than the prevalence of 68% at week 16 found by Beaulac and Schwean-Lardner (2018). In contrast, Leishman et al. (2021) reported a mean prevalence of 38% from assessments conducted by farmers who used the same self-assessment protocol, but without prior training and lacking information on reliability. In comparison to turkeys, a lower mean prevalence was found in broilers at the end of the fattening period (47%), which is in the range of reported mean prevalences from 21% to 52% (Dawkins et al., 2017; Göransson et al., 2020). Tahamtani et al. (2018) found lower values of only 10%. They attributed the low prevalence to warm weather conditions during the assessment time and the mitigating effect of the Danish footpad dermatitis monitoring at slaughterhouses. Different weather conditions may also have affected the current study results on



footpad dermatitis, as assessments were implemented both during the warmer and the colder seasons. However, the visits were nearly evenly distributed, with 10 visits on broiler farms and 11 on turkey farms between May and October and 12 and 10 visits during November to April. The alarm values of the evaluation framework for broilers in their last week of life (24% total/6% severe footpad dermatitis) are clearly lower than the German thresholds for footpad dermatitis at the slaughterhouse (40% superficial/20% severe footpad dermatitis) specified in the Implementing Provisions for the Farm Animal Welfare Regulation (Implementing Provisions TierSchNutzV, 2014). This allows the self-assessment to provide an early warning before levels are reached that lead to official sanctions.

The rates of birds dead on arrival and condemnations were also more often in the alarm area in broiler flocks than in turkey flocks. The alarm values for broilers were set more stringent than those for turkeys. At the same time, the broiler flocks reached slightly lower mean prevalences than the turkey flocks (dead on arrival: 0.07% vs. 0.11% and condemnations: 0.99% vs. 1.1%). The Implementing Provisions for the Farm Animal Welfare Regulation (Implementing Provisions TierSchNutzV, 2014) stipulates a threshold for broilers dead on arrival at the slaughterhouse of 0.5% which is higher than the alarm value of the evaluation framework (0.3%). Again, in this way, self-assessment can inform about a welfare issue before official thresholds are in danger of being exceeded. The rates of condemnations and animals dead on arrival in our study were generally lower than those reported in the literature of the last years: for dead on arrival, 0.08% to 0.26% for broilers (Averós et al., 2020; Allen et al., 2023) and 0.3% for turkeys (Marchewka et al., 2020). However, the high rate for broilers was recorded in Spain, with higher risks of heat stress than in Germany. Forseth et al. (2023) reported condemnation rates of 0.7% to 2.2%, depending on slow- or fast-growing genetics which aligns with our result of approximately 1.0%. We found only slightly higher rates in turkeys which were lower than the mean condemnation rates from the literature with 2.4% for female and 7.7% for male birds (Marchewka et al., 2020; Blomvall et al., 2023).

Lameness, particularly in fast-growing broilers and turkeys, is another highly debated welfare issue. Comparing the mean prevalences found for broilers and turkeys, both were relatively low with 4.7% in broilers and 0.2% in turkeys at the end of the fattening period. Here, the target (0.0%) and alarm (4.0%) values of the evaluation framework are identical for both species. Consequently, no turkey farm fell within the alarm area, while 45% of broiler farms did. In the literature, prevalences of up to 25% of lame broilers were reported, 19% in Granquist et al. (2019), 25% in Kittelsen et al. (2017), and 24% in Marchewka et al. (2013). The lower prevalences observed in our study compared to the literature may be due to breeding advances in recent years (Hartcher and Lum, 2019) and the inclusion of slow-growing birds from organic farms in our sample, which are known to have generally better gait scores (Rayner et al., 2020). For turkeys, prevalences as low as 1% to 2% (Ferrante et al., 2019) or even 0.06% (Marchewka et al., 2020) have been reported, indicating a trend toward better welfare in terms of lameness, which is similar to the findings observed on the farms included in our study.

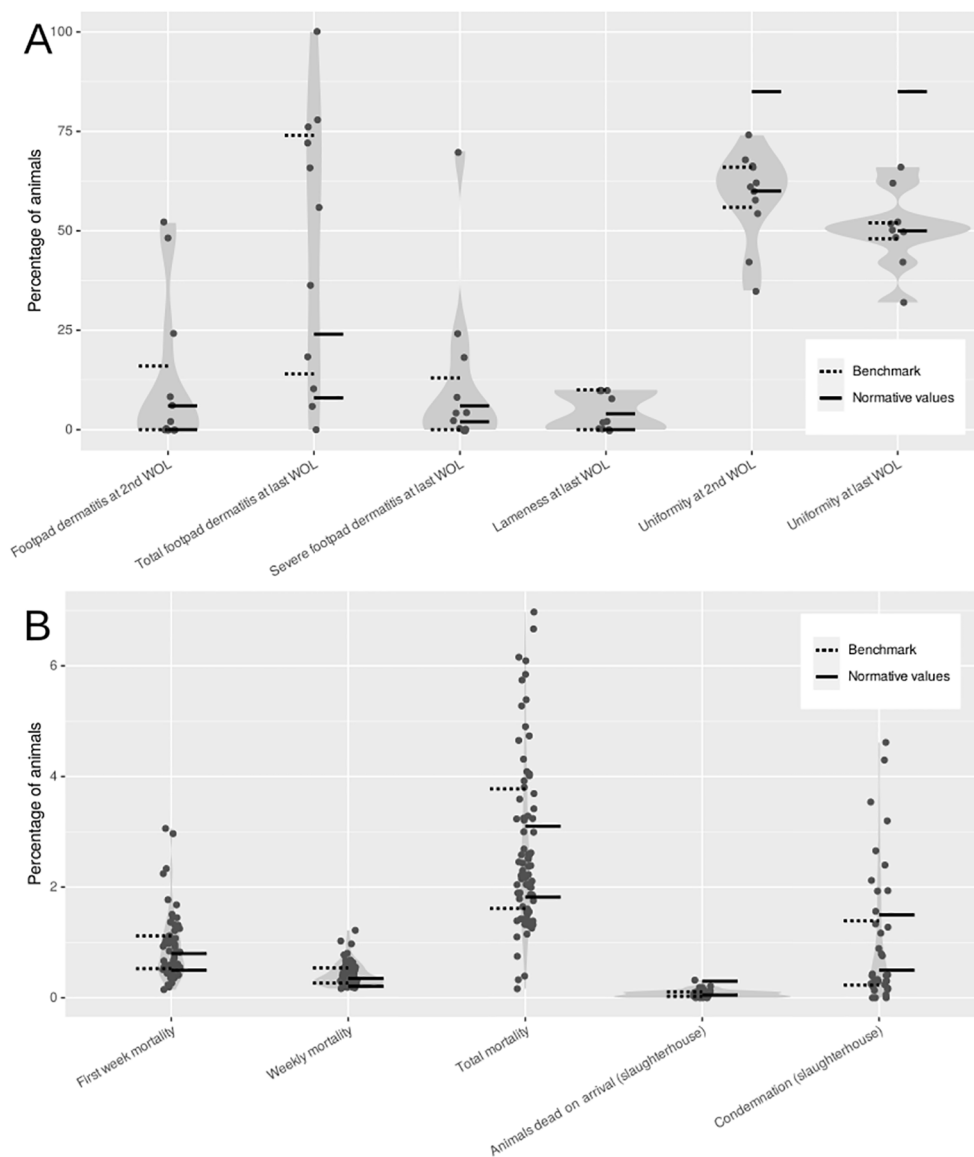
For body weight uniformity, broiler and turkey farms did not reach the target range of at least 85% (broilers) or 90% (turkeys), except for one turkey farm. The average weight uniformity in this study was 50% for broilers and 78% for turkeys. This difference between broilers and turkeys can be explained by the mixed-sex flocks on all broiler farms compared to only three mixed-sex flocks in turkeys. In mixed-sex flocks, reduced weight uniformity can be expected due to the sexual dimorphism in body weight (Gous, 2018). This could be overcome by weighing female and male birds separately, but this requires more effort and is not feasible when automatic weighing is used. Weight uniformity was included as a welfare indicator because low uniformity can reflect the difficulty of lighter and smaller birds in a flock to reach the feeder and drinker lines. Therefore, the alarm values have been set at a relatively low level (50% in broilers, 70% in turkeys at the end of fattening). Still, 56% of the broiler flocks and 30% of turkey flocks fell in the alarm range. In the literature, uniformity in broilers is mostly calculated as coefficient of variation. To the best of our knowledge, no percentages for broilers were reported comparable to our calculations; however, low body weight uniformity has been reported as an issue in broilers (Vasdal et al., 2019; Göransson et al., 2020; Rubio et al., 2023). For turkeys, Jhetam et al. (2022) reported uniformity between 77% and 81% for females and Beaulac and Schwean-Lardner (2018) reported 85% to 89% for males, which is slightly higher than our findings. Since almost none of the farms in our study were in the target range, this suggests that the values may be too strict and may demotivate farmers in their efforts to improve. The participating broiler flocks reached a 25th percentile of 66% in the second week and 52% in the last week of life compared to the target value of 85%.

Regarding total mortality, the evaluations indicated a trend toward better welfare in turkeys, with 41% of broiler flocks in the alarm range compared to just 2% for turkeys. The mean mortality rates for broilers found in this study (3.2%) were in the range of recent literature reports of 2.1% to 3.8% (Tahamtani et al., 2018; BenSassi et al., 2019; Vasdal et al., 2019; Göransson et al., 2020; Dawkins et al., 2021). The alarm value is slightly stricter than the threshold set by the EU Directive, 2007/43/EC, again in order to alert farmers to the risk of exceeding legal limits. Warning and alarm values used in the Welfare Quality<sup>®</sup> protocol for broilers (Welfare Quality<sup>®</sup> Consortium, 2009) are 3% and 6% (when less than 20% of losses are due to culling). These values are nearly twice as high as the values of the evaluation framework (assuming, e.g., 40 days of life). However, the Welfare Quality<sup>®</sup> alarm value would be above the threshold of the EU Directive, 2007/43/EC of 3.4% in 40-day-old broilers. Mortality rates in broilers in this study showed a large between-flock variation, with 35% of the flocks in the alarm range and 35% in the target range. In turkeys, the higher mortality in male compared to female turkeys was confirmed in the test flocks. Only data from 7 female flocks were available, compared to 22 and 26 male flocks for weekly and total mortality rates, respectively. However, on average, 2.1% in female and 5.7% in male turkeys were in line with the reported 3.3% to 3.4% for female birds (Marchewka et al., 2020; Blomvall et al., 2023) and 5.1% to 7.2% for males (Olschewsky et al., 2021; Blomvall et al., 2023). The alarm value for total mortality in female turkeys was set at 0.35%

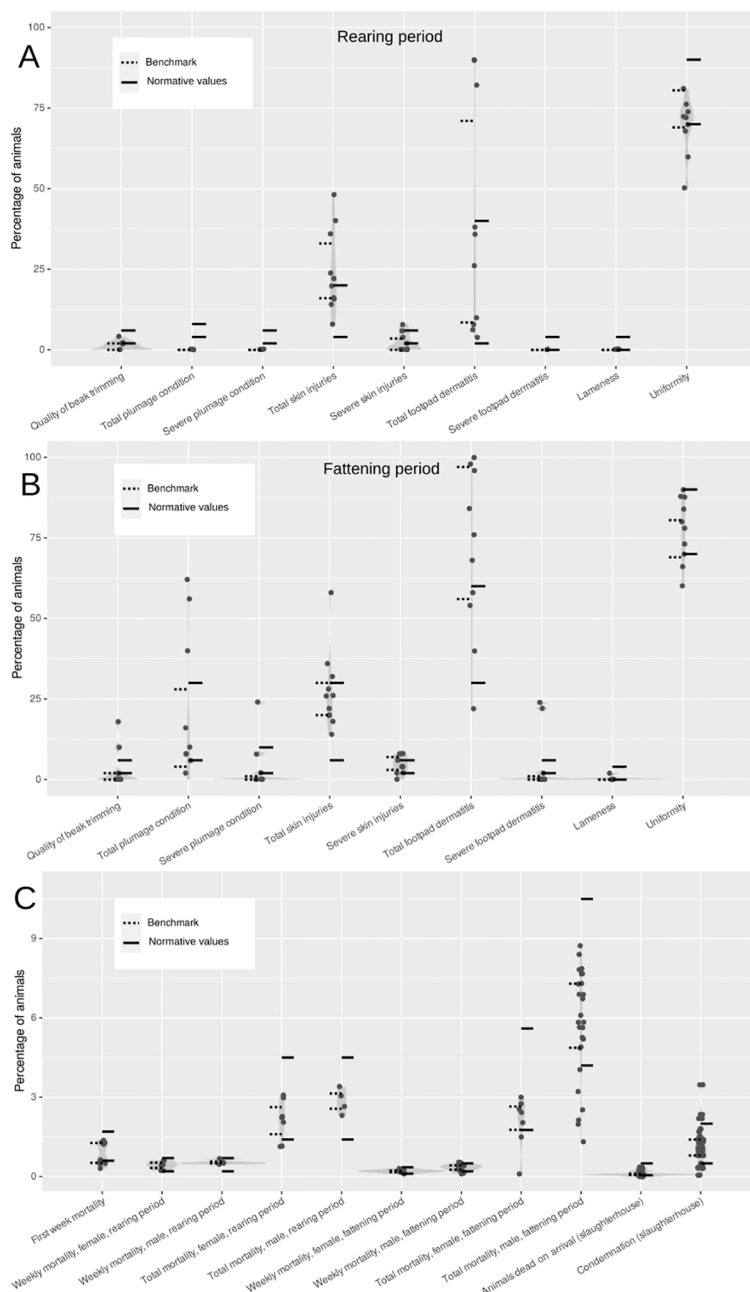
mortality per week, matching that of broilers. However, a higher value of 0.5% was set for male turkeys. For the first-week mortality, the alarm value differed between turkeys (1.7%) and broilers (0.9%), leading again to more broiler flocks (50%) in the alarm area than turkey flocks. The average first-week mortality in broilers in our study (0.93%) was comparable to the rates in the recent literature of 0.01% to 2.02% (Vasdal et al., 2019; Yerpes et al., 2020; Jessen et al., 2021; Yerpes et al., 2021).

The assessments of plumage condition and skin injuries were only performed on turkeys. The alarm values for total alteration (score 1 + 2) in the fattening period were set at a relatively high level (30%) in order to reflect the high prevalences observed in the farm data used in the expert discussions. Despite this, almost a third of

the flocks were in the alarm range for both indicators, indicating a fundamental problem in turkey farming. None of the flocks reached the target area for skin injuries. The mean prevalences reached in this study were 19% for plumage damage and 27% for skin injuries, with zero to low average prevalences in the rearing period. The prevalence of injuries was slightly higher than reported in the literature with 8% in the rearing period (Bartels et al., 2009) and 4% to 23% at the end of fattening (Bartels et al., 2009; Blomvall et al., 2023). Regarding plumage condition, a high variance can be found in the recent literature, with low prevalences of 0% to 1.2% of birds with plumage damage for BUT 6 and Auburn (Grün et al., 2021) to 4% for Kelly bronze and two Hockenhull strains (Olschewsky et al., 2021) and up to 71% for BUT 6 (Haug et al., 2023). It should be



**FIGURE 1** Comparison of normative values of the evaluation framework and the 25th and 75th percentiles as benchmarks for every indicator for broilers. Violin plots represent the prevalences or rates of the provided farm data of the farmers. The violin shape represents the frequency of prevalences or rates. **(A)** The indicators measured on a sample of 50 individual animals; **(B)** the indicators based on flock data. Related *n*-values on the flock level are presented in Table 3. WOL denotes week of life of the broilers.



**FIGURE 2** Comparison of normative values of the evaluation framework and the 25th and 75th percentiles as benchmarks for every indicator for turkeys. Violin plots represent the prevalences or rates of the provided farm data of the farmers. The violin shape represents the frequency of prevalences or rates. **(A)** The indicators measured on a sample of 50 individual animals in the rearing period and **(B)** of the fattening period. **(C)** The indicators based on flock data. Related *n*-values on the flock level are presented in [Table 4](#).

noted that our study recorded prevalences under commercial conditions (mostly with BUT 6 turkeys), whereas all cited studies kept the turkeys in small groups.

The comparison between the two evaluation methods showed also different results for broilers and turkeys. For broilers, the application of normative values resulted in a significantly lower average welfare rank compared to the benchmarking, although with only small numerical differences, while for turkeys, no difference was found. The latter result was surprising since benchmarks

potentially underestimate welfare issues that occur frequently at farms (Bergschmidt et al., 2021), such as for footpad dermatitis. The same reasons discussed earlier for the differences between broiler and turkey evaluations are likely applicable here: the evaluation framework for broilers had stricter limits than for turkeys. Nevertheless, the benchmark approach showed examples of underestimation of welfare problems, very clearly visible for footpad dermatitis for both species (Figures 1, 2). Flocks in the best quartile might be misinterpreted as having a good score, even

though the prevalences are relatively high. This was also the case for body weight uniformity (here low percentages) and total skin injuries. On the other hand, for total skin injuries, the normative evaluation resulted in no flock in the target area, which may indicate that skin injury prevalences under 6% are hardly achievable with current farming practices. It may also be demotivating for farmers to apply target values that are far away from the welfare status of their own birds. However, given that skin injuries cause pain (Yoshiyama et al., 2021), ethical values that prioritize early intervention are more important than motivational factors. The results from the turkey flocks also showed that for certain indicators, the benchmarking led to a more rigorous evaluation. For example, for mortality, no flock fell into the alarm range based on normative values. However, for flocks in the worst quartile of the benchmark, the evaluation indicated poor performance requiring action, even though these farms were within the normative early warning or target range. In such cases, it is necessary to decide whether normative values need to be adjusted or whether further improvement is not urgent because the results are already satisfactory.

Altogether, the comparison between the two evaluation methods confirmed both their advantages and disadvantages. Benchmarks have the advantage of not requiring the initial effort of gathering expert opinions on normative values (Kaurivi et al., 2020; Sapkota et al., 2022) along with the challenges of reaching a consensus. A key element of a benchmark is the visibility of peer-farmers' performance, which can serve as inspiration or motivation (Sumner et al., 2018). However, the resulting percentiles depend on the sample population and are more robust with a larger number of participating farms. This requires a certain infrastructure for collecting, storing, and processing farm data and finally reporting to participants. Additionally, this service needs to be continuously provided since benchmarks need regular updates. For example, in the private sector, turkey farmers can participate in the German Turkey Health Control Program (Andersson et al., 2016). This program provides regular benchmarks based on data collected at the slaughterhouse, such as mortality rates and footpad dermatitis (Toppel et al., 2019). If a farm is classified in the worst quarter, an action plan with the farm veterinarian is required. Another example is the monitoring of antibiotics by a centralized governmental benchmark which reduced the use of antibiotics in Germany over several years successfully (Federal Institute for Risk Assessment, 2021). However, such benchmarks are costly and require a complex reporting system (BearingPoint GmbH, 2021). To date, for Germany, nationwide benchmarks on animal welfare are not available due to the absence of a harmonized collection of data and central databases (Bergschmidt et al., 2023; Johns et al., 2023). This also applies to the European Union (EFSA, 2012). Therefore, farmers wanting to take part in benchmarking require access to private infrastructure. Examples of normative values in the broiler sector include the evaluation framework developed for the Welfare Quality<sup>®</sup> protocol which is mainly used in a scientific context and the legally binding thresholds of the EU Directive, 2007/43/EC for the protection of broilers. However, to the best of our knowledge, there was a lack of normative values to support farmers in their daily work. Normative values can be used independently by any farmer, including those who are not vertically integrated. The evaluations can be implemented at self-selected time intervals, e.g.,

already before depopulation, to take advantage of early warnings. In a number of cases, normative values will lead to stricter evaluations compared to benchmarks, especially where farm results outside the worst quartile can still raise serious societal concerns. However, this scenario may also arise for normative values that take into account the status quo in practice. Both evaluation methods can also be complementary: on the one hand, benchmarking can identify common challenges (e.g., due to seasonal influences) or advances, and on the other hand, the application of normative values can verify that ethical standards are taken into account.

## 5 Conclusion

The application of an evaluation framework for the self-assessment of animal welfare with normative target and alarm values showed that high percentages of flocks were evaluated as being in the alarm range for indicators such as footpad dermatitis or weight uniformity in broilers and turkeys and plumage condition and skin injuries in turkeys. Our hypothesis was confirmed for broilers, as the normative evaluation was significantly stricter than the benchmark across all indicators and flocks. However, no difference could be found for turkeys. Possible reasons for the differences between broilers and turkeys highlight general challenges of the evaluation systems: the farm selection process probably favored turkey farms with better than average management, pointing to the need for a robust, preferably large sample for benchmarking. At the same time, the normative values made allowances for a poorer status quo in turkeys for some indicators, pointing at the challenges of agreeing to normative values in a participatory process. Therefore, normative values may have advantages where large databases for benchmarking are lacking, and they may lead to a more ambitious action where large proportions of animals are affected by welfare problems. They allow evaluations to be made at any chosen point in time, while benchmarking allows comparison with peer farmers where the necessary infrastructure is in place. Normative values should be periodically re-examined, for example, by comparing them with farm data percentiles. Additionally, the latest scientific evidence should be incorporated into this process.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The animal studies were approved by the Designated Veterinarian for Institutional Animal Care of the University of Kassel. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was obtained from the owners for the participation of their animals in this study.

## Author contributions

SM: Conceptualization, Data curation, Formal analysis, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. DG: Conceptualization, Investigation, Writing – review & editing. UK: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study was part of the larger project entitled “Feasibility of Animal Welfare Indicators for On-Farm Self-Assessment: Development of an Orientation Framework With Reference Values and Technical Implementation in Digital Applications (EiKoTiGer),” conducted at the University of Kassel, Germany. This project received funding from the German Federal Ministry of Food and Agriculture (BMEL) pursuant to a decision of the Parliament of the Federal Republic of Germany and administered by the Federal Office for Agriculture and Food ([www.ble.de/ptble/innovationsfoerderung-bmel](http://www.ble.de/ptble/innovationsfoerderung-bmel); grant no. 2817901115) under the Innovation Scheme. The funder played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

- Allain, V., Huonnic, D., Rouina, M., and Michel, V. (2013). Prevalence of skin lesions in Turkeys at slaughter. *Br. Poultry Sci.* 54, 33–41. doi: 10.1080/00071668.2013.764397
- Allen, S., Parker, C., Verheyen, K., Nicol, C., and Chang, Y. (2023). Effects of external ambient temperature at loading, journey duration and flock characteristics on the dead-on-arrival rate in broiler chickens transported to slaughter in Great Britain. *Poultry Sci.* 102, 102634. doi: 10.1016/j.psj.2023.102634
- Andersson, R., Toppel, K., and Heesen, S. (2016). “Kann man Tierwohl messen?,” in *Geflügeljahrbuch 2016*. Eds. K. Damme and F. Muth (Ulmer Verlag, Stuttgart), 24–32.
- Atkinson, D., von Keyserlingk, M., and Weary, D. (2017). Benchmarking passive transfer of immunity and growth in dairy calves. *J. Dairy Sci.* 100, 3773–3782. doi: 10.3168/jds.2016-11800
- Averós, X., Balderas, B., Cameno, E., and Estevez, I. (2020). The value of a retrospective analysis of slaughter records for the welfare of broiler chickens. *Poultry Sci.* 99, 5222–5232. doi: 10.1016/j.psj.2020.08.026
- Bartels, T., Böhme, J., Cramer, K., Dellavolpe, A., Ellerich, R., Ludewig, M., et al. (2009). *Abschlussbericht zum Forschungsauftrag 06HS015 - Indikatoren einer tiergerechten Mastputenhaltung*. Available online at: [https://www.ml.niedersachsen.de/download/72908/Abschlussbericht\\_zum\\_Forschungsauftrag\\_06HS015\\_Indikatoren\\_einer\\_tiergerechten\\_Mastputenhaltung.pdf](https://www.ml.niedersachsen.de/download/72908/Abschlussbericht_zum_Forschungsauftrag_06HS015_Indikatoren_einer_tiergerechten_Mastputenhaltung.pdf) (accessed May 4, 2024).
- BearingPoint GmbH (2021). *Studie zur Antibiotikaverbrauchsmengenerfassung bei Tieren gemäß der Verordnung (eu) 2019/6 - Abschlussbericht*. Available online at: [https://www.bmel.de/SharedDocs/Downloads/DE/\\_Tiere/Tiergesundheit/Tierarzneimittel/studie-antibiotikaverbrauch-tiere-eu-2019-6-lang.pdf?\\_\\_blob=publicationFile&v=2](https://www.bmel.de/SharedDocs/Downloads/DE/_Tiere/Tiergesundheit/Tierarzneimittel/studie-antibiotikaverbrauch-tiere-eu-2019-6-lang.pdf?__blob=publicationFile&v=2) (accessed May 4, 2024).
- Beaulac, K., and Schwan-Lardner, K. (2018). Assessing the effects of stocking density on Turkey tom health and welfare to 16 weeks of age. *Front. Veterinary Sci.* 5, 213. doi: 10.3389/fvets.2018.00213
- BenSassi, N., Vas, J., Vasdal, G., Averós, X., Estévez, I., and Newberry, R. C. (2019). On-farm broiler chicken welfare assessment using transect sampling reflects environmental inputs and production outcomes. *PLoS One* 14, e0214070. doi: 10.1371/journal.pone.0214070
- Bergschmidt, A., Andersson, R., Bielicke, M., Brinkmann, J., Gröner, C., Heil, N., et al. (2023). *Empfehlungen für die Einführung eines nationalen Tierwohl-Monitorings:*

## Acknowledgments

We would like to express our gratitude to the farmers and personnel of the poultry farms for participating in this study, especially for their submitted farm data and their effort during the farm visits. We extend our gratitude to the participants of the Delphi survey and expert discussions for their feedback. Finally, we would like to thank Carlo Michaelis (University of Göttingen, Third Institute of Physics) for his valuable assistance with the statistical analyses.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*6 Punkte zur Umsetzung* (Braunschweig, Germany: Johann Heinrich von Thünen-Institut). doi: 10.3220/MX1686754159000

Bergschmidt, A., March, S., Wagner, K., and Brinkmann, J. (2021). A results-oriented approach for the animal welfare measure of the European Union's rural development programme. *Animals* 11, 1570. doi: 10.3390/ani11061570

Blomvall, L., Kaukonen, E., Kurittu, P., Heikinheimo, A., and Fredriksson-Ahomaa, M. (2023). Food chain information and post-mortem findings in fattening Turkey flocks. *Food Control* 150, 109739. doi: 10.1016/j.foodcont.2023.109739

Botreau, R., Veissier, I., and Pernel, P. (2009). Overall assessment of animal welfare: strategy adopted in Welfare Quality®. *Anim. Welfare* 18, 363–370. doi: 10.1017/s0962728600000762

Dawkins, M. S., Roberts, S. J., Cain, R. J., Nickson, T., and Donnelly, C. A. (2017). Early warning of footpad dermatitis and hockburn in broiler chicken flocks using optical flow, bodyweight and water consumption. *Veterinary Rec.* 180, 499–499. doi: 10.1136/vr.104066

Dawkins, M. S., Wang, L., Ellwood, S. A., Roberts, S. J., and Gebhardt-Henrich, S. G. (2021). Optical flow, behaviour and broiler chicken welfare in the UK and Switzerland. *Appl. Anim. Behav. Sci.* 234, 105180. doi: 10.1016/j.applanim.2020.105180

EFSA (2012). Statement on the use of animal-based measures to assess the welfare of animals. *EFSA J.* 10, 2767. doi: 10.2903/j.efsa.2012.2767

Estevez, I., Marchewka, J., Ferrante, V., and Watanabe, T. (2015). *AWIN welfare assessment protocol for Turkeys*. doi: 10.13130/AWIN.TURKEYS.2015

EU Directive (2007). *Council directive 2007/43/ec, laying down minimum rules for the protection of chickens kept for meat production*.

Federal Institute for Risk Assessment (2021). *Therapiehäufigkeiten und Antibiotikaverbrauchsmengen - Wissenschaftsbericht*. Available online at: <https://www.bfr.bund.de/cm/343/therapiehaeufigkeit-und-antibiotika-verbrauchsmengen-2022-entwicklung-in-zur-fleischerzeugung-gehaltenen-rindern-schweinen-huehner-und-puten.pdf> (accessed May 4, 2024).

Ferrante, V., Lolli, S., Ferrari, L., Watanabe, T. T. N., Tremolada, C., Marchewka, J., et al. (2019). Differences in prevalence of welfare indicators in male and female Turkey flocks (meleagris gallopavo). *Poultry Sci.* 98, 1568–1574. doi: 10.3382/ps/pey534

- Forseth, M., Moe, R., Kittelsen, K., Skjerve, E., and Toftaker, I. (2023). Comparison of carcass condemnation causes in two broiler hybrids differing in growth rates. *Sci. Rep.* 13, 4195. doi: 10.1038/s41598-023-31422-0
- Fraser, D. (1995). Science, values and animal welfare: Exploring the 'inextricable connection'. *Anim. Welfare* 4, 103–117. doi: 10.1017/s0962728600017516
- Freihold, D., Bartels, T., Bergmann, S., Berk, J., Deerberg, F., Dressel, A., et al. (2019). Investigation of the prevalence and severity of foot pad dermatitis at the slaughterhouse in fattening Turkeys reared in organic production systems in Germany. *Poultry Sci.* 98, 1559–1567. doi: 10.3382/ps/pey473
- Göransson, L., Yngvesson, J., and Gunnarsson, S. (2020). Bird health, housing and management routines on Swedish organic broiler chicken farms. *Animals* 10, 2098. doi: 10.3390/ani10112098
- Gous, R. (2018). Nutritional and environmental effects on broiler uniformity. *World's Poultry Sci. J.* 74, 21–34. doi: 10.1017/s0043933917001039
- Granquist, E., Vasdal, G., de Jong, I., and Moe, R. (2019). Lameness and its relationship with health and production measures in broiler chickens. *Animal* 13, 2365–2372. doi: 10.1017/s1751731119000466
- Grün, S., Damme, K., Müller, M., Sommer, M. F., Schmidt, P., Erhard, M., et al. (2021). Welfare and performance of three Turkey breeds—comparison between infrared beak treatment and natural beak abrasion by pecking on a screed grinding wheel. *Animals* 11, 2395. doi: 10.3390/ani11082395
- Hartcher, K., and Lum, H. (2019). Genetic selection of broilers and welfare consequences: a review. *World's Poultry Sci. J.* 76, 154–167. doi: 10.1080/00439339.2019.1680025
- Haug, D., Schreiter, R., Thesing, B., Rathmann, L., Lambertz, C., Hofmann, P., et al. (2023). Injurious pecking in organic Turkey fattening – effects of husbandry and feeding on injuries and plumage damage of a slow- (auburn) and a fast-growing (B.U.T.6) genotype. *Poultry Sci.* 102, 102746. doi: 10.1016/j
- Heidemark (2023). *Viel.mehr.nachhaltigkeit - Nachhaltigkeitsbericht 2023*. Available online at: <https://www.yumpu.com/de/document/read/68471392/vielmehrernachhaltigkeit-nachhaltigkeitsbericht-2023> (accessed May 4, 2024).
- Hübel, J. (2019). *Fußballentzündung, Einstreufeuchtigkeit und Mortalität als Tierschutzindikatoren in der Aufzuchtphase von Mastputen unter Berücksichtigung von Besatzdichte und Körpermasse*. Ph.D. thesis. (Leipzig, Germany: University Leipzig).
- Implementing Provisions TierSchNutzV (2014). *Ausführungsbestimmungen zur Umsetzung von § 20 abs. 4 und 5 TierSchNutzV [implementing provisions for § 20 (4) and (5) of the animal welfare livestock husbandry ordinance]. As of 22 August 2006 (BGBl. I p. 2043), last amended 1 October 2009. 3223.*
- Jessen, C. T., Foldager, L., and Riber, A. B. (2021). Effects of hatching on-farm on performance and welfare of organic broilers. *Poultry Sci.* 100, 101292. doi: 10.1016/j.psj.2021.101292
- Jhetam, S., Buchynski, K., Shynkaruk, T., and Schwan-Lardner, K. (2022). Stocking density effects on Turkey hen performance to 11 weeks of age. *Poultry Sci.* 101, 101874. doi: 10.1016/j.psj.2022.101874
- Johns, J., Andersson, R., Bielick, M., Brinkmann, J., Fritzen, D., Gröner, C., et al. (2023). *Abschlussbericht Nationales Tierwohlmonitoring (NaTiMon)*.
- Kaurivi, B., Hickson, R., Laven, R., Parkinson, T., and Stafford, K. (2020). Developing an animal welfare assessment protocol for cows in extensive beef cow-calf systems in New Zealand. part 2: categorisation and scoring of welfare assessment measures. *Animals* 10, 1592. doi: 10.3390/ani10091592
- Kittelsen, K., David, B., Moe, R., Poulsen, H., Young, J., and Granquist, E. (2017). Associations among gait score, production data, abattoir registrations, and postmortem tibia measurements in broiler chickens. *Poultry Sci.* 96, 1033–1040. doi: 10.3382/ps/pew433
- Knierim, U., Andersson, R., Keppler, C., Petermann, S., Rauch, E., Spindler, B., et al. (2016). *Tierschutzindikatoren: Leitfaden für die Praxis - Geflügel* (Darmstadt: Kuratorium für Technik und Bauwesen in der Landwirtschaft).
- Knierim, U., Carter, C., Fraser, D., Gärtner, K., Lutgendorf, S., Mineka, S., et al. (2001). "Good welfare: improving quality of life," in *Coping with challenge: Welfare in animals including humans, Dahlem Workshop Report 87*, 79–100.
- Knierim, U., Gieseke, D., Michaelis, S., Keppler, C., Spindler, B., Rauch, E., et al. (2020a). *Tierschutzindikatoren: Leitfaden für die Praxis - Geflügel. 2nd edn* (Darmstadt: Kuratorium für Technik und Bauwesen in der Landwirtschaft). Available online at: <https://www.ktbl.de/themen/tierwohlbewertung>.
- Knierim, U., Michaelis, S., Gieseke, D., Schultheiß, U., and Zapf, R. (2020b). *Tierschutzindikatoren für Masthühner: Vorschläge zu Ziel- und Alarmwerten für die betriebliche Eigenkontrolle*. Available online at: [https://www.ktbl.de/fileadmin/user\\_upload/Allgemeines/Download/Tierwohl/KTBL-UniKassel-Tierschutzindikatoren\\_Ziel-undAlarmwerte\\_Masthuehner.pdf](https://www.ktbl.de/fileadmin/user_upload/Allgemeines/Download/Tierwohl/KTBL-UniKassel-Tierschutzindikatoren_Ziel-undAlarmwerte_Masthuehner.pdf) (accessed May 4, 2024).
- Knierim, U., Michaelis, S., Gieseke, D., Schultheiß, U., and Zapf, R. (2020c). *Tierschutzindikatoren für Mastputen: Vorschläge zu Ziel- und Alarmwerten für die betriebliche Eigenkontrolle*. Available online at: [https://www.ktbl.de/fileadmin/user\\_upload/Allgemeines/Download/Tierwohl/KTBL-UniKassel-Tierschutzindikatoren\\_Ziel-undAlarmwerte\\_Mastputen.pdf](https://www.ktbl.de/fileadmin/user_upload/Allgemeines/Download/Tierwohl/KTBL-UniKassel-Tierschutzindikatoren_Ziel-undAlarmwerte_Mastputen.pdf) (accessed May 4, 2024).
- Knierim, U., and Winckler, C. (2009). On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. *Anim. Welfare* 18, 451–458. doi: 10.1017/s0962728600000865
- Leishman, E. M., van Staaveren, N., Osborne, V. R., Wood, B. J., Baes, C. F., and Harlander-Matuschek, A. (2021). A cross-sectional study on the prevalence of footpad dermatitis in Canadian Turkeys. *Front. Anim. Sci.* 2, 726907. doi: 10.3389/fanim.2021.726907
- Main, D. C. J., Kent, J. P., Wemelsfelder, F., Ofner, E., and Tuytens, F. A. M. (2003). Applications for methods of on-farm welfare assessment. *Anim. Welfare* 12, 523–528. doi: 10.1017/s0962728600026129
- Marchewka, J., Vasdal, G., and Moe, R. (2020). Associations between on-farm welfare measures and slaughterhouse data in commercial flocks of Turkey hens (mealegris gallopavo). *Poultry Sci.* 99, 4123–4131. doi: 10.1016/j.psj.2020.05.036
- Marchewka, J., Watanabe, T., Ferrante, V., and Estevez, I. (2013). Welfare assessment in broiler farms: transect walks versus individual scoring. *Poultry Sci.* 92, 2588–2599. doi: 10.3382/ps.2013-03229
- Michaelis, S., Schubert, A., Gieseke, D., Cimer, K., Zapf, R., Lühken, S., et al. (2022). A comparison of online and live training of livestock farmers for an on-farm self-assessment of animal welfare. *Front. Anim. Sci.* 3, 915708. doi: 10.3389/fanim.2022.915708
- Olschewsky, A. (2019). *Untersuchung der Eignung alternativer Putenherkünfte für ein ökologisches Haltungssystem*. Ph.D. thesis (Kassel, Germany: University Kassel).
- Olschewsky, A., Riehn, K., and Knierim, U. (2021). Suitability of slower growing commercial Turkey strains for organic husbandry in terms of animal welfare and performance. *Front. Veterinary Sci.* 7, 600846. doi: 10.3389/fvets.2020.600846
- Pandolfi, F., Stoddart, K., Wainwright, N., Kyriazakis, I., and Edwards, S. (2017). The 'real welfare' scheme: benchmarking welfare outcomes for commercially farmed pigs. *Animal* 11, 1816–1824. doi: 10.1017/s1751731117000246
- Rayner, A. C., Newberry, R. C., Vas, J., and Mullan, S. (2020). Slow-growing broilers are healthier and express more behavioural indicators of positive welfare. *Sci. Rep.* 10, 15151. doi: 10.1038/s41598-020-72198-x
- Rösler, B. (2016). *Untersuchungen von konventionell gehaltenen Ross 308 Masthühnern in einer angereicherten Haltungsumwelt unter dem Aspekt der Tiergesundheit*. Ph.D. thesis. (Munich, Germany: Ludwig-Maximilians-University Munich).
- Rubio, A., Conrad, S., Juzaitis-Boelter, C., Wishon, C., and Fahrenholz, A. (2023). The impact of marker selection, in-line near-infrared spectroscopy (nir), and feed mix time on the coefficient of variation (mix uniformity), body weight uniformity and broiler growth performance during the starter, grower, and finisher periods. *Poultry Sci.* 102, 103109. doi: 10.1016/j.psj.2023.103109
- Sanders, P., Vanderhaeghen, W., Fertner, M., Fuchs, K., Obritzhauser, W., Agunos, A., et al. (2020). Monitoring of farm-level antimicrobial use to guide stewardship: overview of existing systems and analysis of key components and processes. *Front. Veterinary Sci.* 7, 540. doi: 10.3389/fvets.2020.00540
- Sapkota, S., Laven, R., Müller, K. R., and Kells, N. (2022). Practicability of a time-limited welfare assessment protocol for pasture-based dairy farms, and a preliminary assessment of welfare outcome thresholds. *Animals* 12, 2481. doi: 10.3390/ani12182481
- Schultheiß, U., Zapf, R., Brinkmann, J., Cimer, K., March, S., Schrader, L., et al. (2023). Tools for on-farm self-assessment of animal welfare. *Landtechnik* 78, 114–123. doi: 10.15150/LT.2023.3293
- Sumner, C. L., von Keyserlingk, M. A. G., and Weary, D. M. (2018). How benchmarking motivates farmers to improve dairy calf management. *J. Dairy Sci.* 101, 3323–3333. doi: 10.3168/jds.2017-13596
- Sumner, C. L., von Keyserlingk, M. A. G., and Weary, D. M. (2020). How benchmarking promotes farmer and veterinarian cooperation to improve calf welfare. *J. Dairy Sci.* 103, 702–713. doi: 10.3168/jds.2019-16338
- Tahamtani, F. M., Hinrichsen, L. K., and Riber, A. B. (2018). Welfare assessment of conventional and organic broilers in Denmark, with emphasis on leg health. *Veterinary Rec.* 183, 192–192. doi: 10.1136/vr.104817
- TierSchG (2006). *Tierschutzgesetz [animal welfare act]. As of 18 May 2006 (BGBl. I p. 1206, 1313), last amended 20 December 2022 (BGBl. I p. 2752).*
- Toppel, K. (2020). *Indicator-based flock management in broilers and Turkeys - mortality, foot pad health and environmental factors*. Ph.D. thesis (Göttingen, Germany: Georg-August University Göttingen).
- Toppel, K., Spindler, B., Kaufmann, F., Gaulty, M., Kemper, N., and Andersson, R. (2019). Foot pad health as part of on-farm-monitoring in Turkey flocks. *Front. Veterinary Sci.* 6, 25. doi: 10.3389/fvets.2019.00025
- Vasdal, G., Granquist, E. G., Skjerve, E., de Jong, I. C., Berg, C., Michel, V., et al. (2019). Associations between carcass weight uniformity and production measures on farm and at slaughter in commercial broiler flocks. *Poultry Sci.* 98, 4261–4268. doi: 10.3382/ps/pez252
- von Keyserlingk, M. A. G., Barrientos, A., Ito, K., Galo, E., and Weary, D. M. (2012). Benchmarking cow comfort on North American freestall dairies: lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. *J. Dairy Sci.* 95, 7399–7408. doi: 10.3168/jds.2012-5807
- Welfare Quality® Consortium (2009). *Welfare Quality® assessment protocol for poultry (broilers, laying hens)* (Lelystad, Netherlands).
- Westermaier, C. (2015). *Vergleichende Untersuchungen zur Tiergesundheit von konventionell gehaltenen Ross 308 und Cobb Sasso Masthühnern mit einem neuen Aufzuchtprozess im Rahmen der konzeptionellen Ausarbeitung von Richtlinien für eine*

*tierrgerechtere Masthühnerhaltung*. Ph.D. thesis (Munich, Germany: Ludwig-Maximilians-University Munich).

World Organisation for Animal Health (2024). *Animal welfare*. Available online at: <https://www.woah.org/en/what-we-do/animal-health-and-welfare/animal-welfare/> (accessed May 4, 2024).

Yerpes, M., Llonch, P., and Manteca, X. (2020). Factors associated with cumulative first-week mortality in broiler chicks. *Animals* 10, 310. doi: 10.3390/ani10020310

Yerpes, M., Llonch, P., and Manteca, X. (2021). Effect of environmental conditions during transport on chick weight loss and mortality. *Poultry Sci.* 100, 129–137. doi: 10.1016/j.psj.2020.10.003

Yoshiyama, Y., Sugiyama, Y., Ishida, K., Fuseya, S., Tanaka, S., and Kawamata, M. (2021). Plantar incision with severe muscle injury can be a cause of long-lasting postsurgical pain in the skin. *Life Sci.* 275, 119389. doi: 10.1016/j.lfs.2021.119389

Zapf, R., Schultheiß, U., Achilles, W., Schrader, L., Knierim, U., Herrmann, H.-J., et al. (2015). Indicators for on-farm self-assessment of animal welfare – example: dairy cows. *Landtechnik* 70, 221–230. doi: 10.1515/LT.2015.2678

Zapf, R., Schultheiß, U., Schubbert, A., Gieseke, D., Cimer, K., March, S., et al. (2023). Evaluation of animal welfare in on-farm self-assessments – development of a reference framework with target and alarm values based on a delphi survey. *Landtechnik* 78, 125–138. doi: 10.1515/LT.2023.3294