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

Animal Welfare 2016, 25: 207-215
ISSN 0962-7286
doi: 10.7120/09627286.25.2.207

Comparison of animal welfare indices in dairy herds based on different sources of data

ND Otten*†, T Rousing‡, H Houe†, PT Thomsen‡ and JT Sørensen‡

† Department of Large Animal Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg C, Denmark
‡ Department of Animal Science, Aarhus University, PO Box 50, DK-8830 Tjele, Denmark
* Contact for correspondence and requests for reprints: nio@sund.ku.dk

Abstract

The present study seeks to evaluate the potential of a more cost-efficient animal welfare assessment by investigating the association between animal welfare indices (AWI) based on different data sources, namely register data (AWI 1, ie routine registrations, such as treatment, reproduction and abattoir data) and resource data (AWI 2, ie barn design and equipment) validated against animal-based data (AWI 3, ie direct animal observations). AWIs were created based on data from 73 Danish dairy herds. Indices for each information source were created by a weighted linear aggregation of herd level incidence and prevalence of the given indicators. Indicator weights were assigned by expert opinion for each of the AWIs. Linear dependency between the high cost AWI 3 and the two low cost AWI 1 and AWI 2 was investigated. Additionally, different time-periods of 90, 180 and 365 days prior to the actual on-farm collection of AWI 3 measures were evaluated in order to find the most predictive time-period of AWI 1. Predictive key indicators for on-farm animal welfare were investigated in uni- and multivariable analyses. Significant associations were found between the AWI 1 based on incidences 180 days prior to the farm visit and the AWI 3. Predictive key indicators were milk yield, abattoir and mortality data. Predictive models for 180 and 365 days prior to the on-farm assessment consisted of abattoir indicators, while the model 90 days prior included mortality and milk yield. The limited associations between indices and the predictive key indicators and models suggest that these cost-effective welfare assessments are not suitable to stand alone and cannot replace the actual animal welfare assessed by on-farm collection of animal-based measures.

Keywords: *aggregation model, animal-based measures, animal welfare, dairy cattle, register data, resource-based measures*

Introduction

The assessment and quantification of animal welfare has been of major concern over previous decades with differing approaches showing different emphasis on measures belonging to different data sources. The animal-based measures are believed to assess animal welfare in the most valid way (Keeling 2009) as these output measures are directly reflecting the animal's response to the input or so-called resource-based measures. Although, resource-based measures are more cost effective due to being easily obtainable and less time-consuming than animal-based measures, studies indicate that animal welfare measured by animal-based measures may vary within the same or similar housing systems and overall management regimes that are alike (Whay *et al* 2003; Rousing *et al* 2007). This challenges the direct comparison of welfare assessment outcomes in different herds and the prospect of reducing the number of animal-based measures in assessment protocols (Mulleder *et al* 2007). To date, the most comprehensive welfare assessment protocol was given by the Welfare Quality® (WQ) assessment protocol in 2009 (Welfare Quality® 2009a,b,c). However, a major drawback for implementation of the WQ protocol is costs due to the major time consumption of data collection. Furthermore, in order to validate the objective nature of the included measures, welfare assessors need to perform ongoing calibration, all adding to the total cost of these animal-based assessment protocols. The estimated time consumption for a full WQ assessment is 7–8 h for a 200 head dairy herd (Welfare Quality® 2009a).

For practical and economic reasons there is a need for a more cost-efficient approach than the WQ. In order to allocate resources in a more beneficial manner, screening tools for animal welfare could be of interest. This could be achieved by using predictive welfare indicators as found in existing databases, ie existing routine registrations made by farmers, inseminators, veterinarians, dairy programmes, laboratories etc, hereafter termed secondary animal-based measures, as register data are already available, however, originally collected and used for other purposes. According to EFSA (2012) this database approach could also facilitate an objective approach for the quantitative risk assessment of animal welfare. However, the knowledge on how well an animal welfare assessment based on these secondary animal-based measures correlates to an animal welfare assessment based on primary animal-based measures is scarce. This possibility has been investigated previously in different settings. In a recent review, deVries *et al* (2011) advocated the possible use of these fairly cheap and readily accessible indicators as an alternative welfare assessment. Although, previous studies investigated either single register data indicators or combinations of indicators for predictive potential in regards to dairy cattle welfare (Sandgren *et al* 2009; Kelly *et al* 2011; Nyman *et al* 2011) and pig welfare (Dewey *et al* 2009; Knage-Rasmussen *et al* 2015), no attempts have been made to make a full register-based quantification of dairy cattle welfare. Despite the differences in the available register-based indicators on a national level, the most frequently seen welfare-associated indicator was mortality both for dairy cow welfare as an overall definition (Sandgren *et al* 2009; Kelly *et al* 2011; Nyman et al 2011) and with the single animal- and resource-based measures within the WQ® protocol (deVries *et al* 2014) followed by fertility measures. Health records, milk production and abattoir data were included in a prediction model for non-compliance issues with animal welfare legislation presented by Otten *et al* (2014). To our knowledge, no previous studies have compared animal welfare based on different information levels. Since register-based indicators are related to several animal-based welfare indicators, a welfare assessment based on these register data could serve as a screening tool for detecting herds with potential welfare problems for subsequent on-farm welfare assessment.

Hence, the overall aim of the present study was to explore the possibility of predicting the actual on-farm dairy welfare without visiting the farm but rather using existing data in the Danish Cattle Database (DCD). This was done by pursuing the following objectives: firstly, to establish animal welfare indices for three different levels of measures with increasing costs from the low-cost, register-based measures, over the easily feasible resource-based measures, to the expensive direct on-farm, animal-based measures. Secondly, to evaluate the associations between the three indices and subsequently establishing the most appropriate time-period for register data to predict the actual on-farm welfare. Finally, key indicators among register-based indicators were identified based on their association with the actual on-farm welfare measured directly on animal outcomes.

Materials and methods

On-farm assessment

The study herds were drawn from a list of 401 responders to a survey on grazing strategies carried out by the Danish Cattle Federation and Aarhus University in 2009. The target population was a Danish dairy herd with more than 100 cows. In order to be representative for the typical Danish dairy herd, loose-housing systems with cubicles were added as additional inclusion criteria for the study herds in the present study. Out of the 401 respondents, two random samples were drawn for two separate studies and hereafter amalgamated for the present study. The first sample of 41 herds came from a study on the effect of grazing on the overall welfare (Burow *et al* 2013). From the same pool of initial respondents a new sample of 90 herds was drawn, of which 45 were willing to participate. Each herd was visited once by one of four trained observers during a period from April 2010 to July 2011. Due to non-compliance with the inclusion criteria upon data collection at the visit, four herds were excluded. A further nine herds were excluded due to missing or incomplete data sheets. Finally, 73 herds remained in the study for further analysis of which 40 implemented summer grazing. Within-herd sampling of cows was carried out as per Burow *et al* (2013), with a minimum of 50 cows per herd yielding an assessment total of 4,647 cows.

Each herd visit started approximately 1 h after morning feeding. Observers were trained via two on-farm sessions as well as one video and picture session prior to conducting the visits. Upon completion of data collection, a second video and picture session was performed to assess inter-observer agreement. For each clinical and behavioural measure, a prevalence-adjusted, bias-adjusted kappa (PABAK; Byrt *et al* 1993) was calculated. PABAK values for inter-observer agreement ranged from 0.25–0.83 while mean levels for each measure ranged from unsatisfactory for hair coat to sufficient and good agreement for the other measures (Landis & Koch 1977).

Measures

Register-based measures

A list of 28 register-based indicators was made reflecting the different aspects of a dairy cow's lifecycle (reproduction, milk production, treatments, mortality and abattoir remarks) and representing the aspects of productivity, health and management related to the given categories of welfare (Table 1). Data were extracted from the Danish Cattle Database for three different time-periods prior to the farm visit date of the given herd: i) 365 days prior; ii) 180 days prior; and iii) 90 days prior to the visit date of the herd in question. All indicators were extracted as within herd-level prevalence or means for the given period.

In order to be used in the animal welfare index (AWI), all indicator values were transformed into categorical values based on the distribution of indicator percentiles among the sample. Descriptive statistics showed a broad range of outcomes for all measures as shown in Otten *et al* (2014); hence, data-driven cut-offs based on percentiles were used. A score 0 was given for indicator values among the 25% best herds within the given indicators: score 1 for indicator values > 25th and < 75th percentiles; score 2 for indicator values ≥ 75th and < 90th percentile; and score 3 for indicator values among the worst 10% in the sample. Since the spread of abattoir remarks (total of nine) between farms was very low, only remarks on lung disorders, liver abscesses, peritonitis, cirrhosis and liver flukes, old fractures and chronic inflammation were kept in the model, leaving 24 indicators for further analysis.

https://doi.org/10.7120/09627286.25.2.207 Published online by Cambridge University Press

Table 1 Median (± SEM) weights identified by expert opinion for welfare indicators and measures from three different information sources (register data from routine registrations, resource-based measures and animal-based measures) used for the calculation of an Animal Welfare Index (AWI) for dairy cattle.

Information level	Indicator or measure		Category	Median weight (± SEM)
Level 1	Lean cows at slaughter†		Feeding	3.0 (± 0.26)
Register-based	Abattoir remarks liver cirrhosis			2.0 (± 0.24)
	Bulk tanks somatic cell count		Health	3.0 (± 0.29)
	Veterinary treatments per 100 cow years‡			3.0 (± 0.27)
	Treatments of locomotor disorder per 100 cow years‡			3.0 (± 0.32)
	Proportion of locomotor disorders			3.0 (± 0.32)
	Proportion of abattoir remarks			3.0 (± 0.33)
	Abattoir remarks:	lung disorders		3.0 (± 0.24)
		liver abscesses		3.0 (± 0.18)
		peritonitis		2.5 (± 0.24)
		liver flukes and cirrhosis		2.0 (± 0.24)
		chronic inflammation		2.0 (± 0.24)
	Cow mortality		Management	5.0 (± 0.25)
	Heifer mortality			4.0 (± 0.22)
	Calf mortality			4.0 (± 0.23)
	Annual average milk yield per cow‡			4.0 (± 0.25)
	Milk yield per location group (1st, 2nd or 3rd and above)			4.0 (± 0.23)
	Standard deviation of milk yield per lactation group			4.0 (± 0.25)
	Age at first calving			3.0 (± 0.24)
	Standard deviation of age at first calving			3.0 (± 0.28)
	Abattoir remarks old fractures			2.0 (± 0.18)
Level 2	Water supply		Feeding	4.0 (± 0.22)
Resource-based	Water cleanliness			3.0 (± 0.20)
	Number of feeding slots			4.0 (± 0.23)
	Occupancy rate bed stalls		Housing	4.0 (± 0.18)
	Bed stall length			3.0 (± 0.18)
	Bed stall width			3.0 (± 0.18)
	Passageways:	width		3.0 (± 0.15)
		skid resistance		3.0 (± 0.15)
		flooring		3.0 (± 0.15)
	Dead ends			3.0 (± 0.15)
	Calving pen size		Health	4.0 (± 0.20)
	Separation of animals			4.0 (± 0.20)
	Sick animals not in sick bay			3.0 (± 0.18)
	Harmful/damaged equipment			4.0 (± 0.15)
	Brushes			3.0 (± 0.22)
	Scraping system			3.0 (± 0.23)
Level 3	Body condition score		Feeding	4.0 (± 0.16)
Animal-based	Hygiene	leg	Housing	2.0 (± 0.13)
		hind		2.0 (± 0.17)
		udder		2.0 (± 0.17)
	Rising behaviour			3.0 (± 0.20)
	Integument alterations	carpus	Health	4.0 (± 0.16)
		tarsus		4.0 (± 0.17)
		body		4.0 (± 0.13)
	Hair coat			2.75 (± 0.15)
	Lameness			4.5 (± 0.38)
	Claw conformation		Management	3.75 (± 0.15)
	Avoidance distance			3.0 (± 0.19)

† Cows with a fat score 1 according to the EU Beef Carcase Classification; ‡ Sum of feeding days of all cows per herd per 365 days.

doi: 10.7120/09627286.25.2.207

Resource-based measures

At each herd visit, a total of 127 resource-based variables were recorded. These included qualitative scoring of, eg cleanliness of water points, the slipperiness of the floor, light and sufficiency of bedding material and quantitative measures of, eg cubicle dimensions, passage width, number of feeding slots etc. The resource-based measures were combined to cover 16 indicators compared to the seven indicators in the WQ in order to establish an intermediate cost index as physical measurements require a limited amount of time. However, as shown in Table 1, the resource measures were reflecting the overall aspects of: feed and water provision; resting area; movement and space; sick pens; and barn equipment. In this step, indicators were evaluated based on current Danish recommendations concerning animal welfare and when applicable also under Danish animal welfare legislation (Danish Act no 520 on Keeping Dairy Cattle and their Offspring). Herds were given a value ranging from zero to one, based on compliance with the given recommendations for the indicator in question. Compliance was given a score 0 while non-compliance resulted in a value of 1, whereas partial compliance was regarded as a fraction of the non-complying measures: n measures yielding non-compliance/total possible. In other words, a herd not fulfilling the recommendations for one passage width out of a total of four passages would receive a value 1/4 = 0.25. Missing values were assigned a non-informative score of 0.5.

Primary animal-based measures

A clinical assessment protocol with ten clinical and two behavioural measures was used to score a random sample of cows consisting of both lactating and dry cows in each herd (sampling strategy according to the WQ® protocol). The clinical protocol was modified from the WQ® protocol to fit Danish settings, excluding a number of measures due to very low prevalence. The remaining measures included: hygiene leg; hygiene hindquarter; hygiene udder; integument alteration on carpus, tarsus and body; claw conformation; body condition score (BCS); lameness; and the avoidance distance (AD). Furthermore, the measures hair coat and rising behaviour were added to the protocol based on their implementation in previous evaluation methods (Rousing *et al* 2007; Thomsen *et al* 2007). Eight of the clinical measures were assessed at graded levels 0–2 (absent, moderate and severe impairment), while claw conformation was assessed on a binary scale (normal or overgrown), BCS on an ordinal scale with quarterly intervals (1–5) and rising behaviour on an ordinal scale ranging from 1–5. The AD was assessed at the beginning of each herd visit and measured in centimetres in 10-cm intervals. Finally, the three latter measures were transformed into the graded scale of normal, moderate or severe impairment as follows: $1.75 > \text{BCS} \leq 2.5$ as 'lean' (moderate impairment) and ≤ 1.75 as 'thin' (severe impairment), rising behaviour was graded as 1–2: 'normal', 3: 'interrupted' (moderate) and 4–5: 'abnormal' (severe), $0.5 \leq \text{AD} < 1$ m as 'sceptical' (moderate) and $\text{AD} \geq 1$ m as 'shy' (severe). All measures were used as mean within-herd prevalence and weighted by their severity in the AWI.

Animal Welfare Index (AWI)

Establishing measures and weights

All register- and resource-based measures included in the WQ® were separated from the current protocol, since they were conflicting with the research question whether an index based on these measures, ie AWI 1 or AWI 2 would be associated with the AWI 3, an index based solely on animal-based measures. Hence, three separate indices were created based on the same aggregation model.

An additive linear aggregation approach directly aggregating measures to individual farm scores expressed as herd Animal Welfare Indexes (AWI) was applied. The individual farm welfare index directly reflected the 'proportion' of animals in the herd with remarks referring to the measures included (specified as prevalence as mentioned in the previous section) (Burow *et al* 2013). This additive linear approach follows a more biologically plausible reasoning where, eg the increase from 10 to 20% lame cows is not judged from an ethical point of view, which would require a non-linear approach as done in the WQ®. However, similar to the WQ®, all measures were assigned to one of the four categories 'feeding', 'housing', 'health' and 'management', with the latter replacing 'appropriate behaviour' as a more descriptive term (Table 1). Furthermore, an expert panel was used to: i) measure weights describing the relative 'impact' of measures; and ii) assess the relative impact of individual measure levels (answering the question: 'what is the relative weight of severe vs moderate level of graded measures as, eg lameness?') — both as regards to measuring the contribution to the welfare index. The herd animal welfare index (AWI) was calculated based on the graded and non-graded welfare measures.

Expert panel opinions

Experts belonging to the fields of research, industry, official welfare control, and animal rights organisations were identified by their contribution or participation in previous dairy cattle welfare projects, while bovine practitioners were recruited based on a balanced distribution among experience (more or less than ten years) and sex. Within the expert panel, 20 out of 32 appointed potential panel participants with expertise in dairy cattle welfare, namely within the field of production advice (2), bovine veterinary practice (5), research (3), industry affiliations (6), official welfare control (3) and animal rights organisations (1) responded to an online questionnaire during December 2012 and January 2013. For each measure the experts were required to provide scores on a five-point scale ranging from 1 (non-important) to 5 (very important). Balanced weighting was ensured in so far as panel participants also had to make sure that the overall mean of all weights was 3 (equal importance). The relative weight of severe vs moderate level of graded measures was formulated as open questions — implying that experts were not limited to any specified range beforehand. The graded (moderate/severe) and non-graded measures were first multiplied by the median weight assigned to the respective measure and graded level by the experts and then summed up to calculate the AWI.

https://doi.org/10.7120/09627286.25.2.207 Published online by Cambridge University Press

Aggregation model for calculation of an Animal Welfare Index (AWI)

A simple additive and weighted index was created for the register- (AWI 1) and resource-based (AWI 2) measures:

$$AWI = \sum_{j=1}^{k} N_j W_j$$

where N denotes the *jth* measure and W the *jth* median weight assigned by the experts.

For the animal-based measures (AWI 3), an extended model was used, taking the graded measures into account as well using the formula below:

$$AWI = \sum_{i=1}^{k} (M_i \, MW_i + S_i) \, W_i + \sum_{j=1}^{k} N_j W_j$$

Modified from Burow *et al* (2013) where M, S and N were the herd's adjusted prevalence of moderate measure levels, severe measure levels and non-level graded measure, respectively. MW and W were the expert panel medians of relative measure level weights and measure weights; i was the prevalence of the individual level graded measure, j the proportion of the individual non-graded measure. For the graded measures k was 9 and for non-graded measures k = 1. Hence, the theoretical maximum score for the AWI 1 based on secondary animal-based measures (AWI 1 365 days, 180 days, 90 days) was 196.5, given that all indicators were amongst the 10% worst herd values (score 3). The theoretical maximum scores for AWI 2 based on resource-based measures (all measures with a score 1 for full non-compliance) was 52 and finally the maximum score was 3,900 for AWI 3, based on primary animal-based measures, given that all measures were at a 100% prevalence for the severe levels. Finally, the two low-cost indices (AWI 1 and AWI 2) were added to an overall index (AWI 1+2) and assessed for association with the high-cost index AWI 3.

Data analysis

All data editing, AWI calculations and statistical analyses were made in R (R Development Core Team 2012) using a general linear model with the lm function to assess the associations (significance level $P = 0.05$) between the different AWIs. Predictive key indicators for on-farm animal welfare expressed by the AWI 3 as well as exploration of associations between predictive models were found using a uni- and multivariable screening with backwards elimination at a 5% significance level, respectively.

Results

Descriptive statistics along with experts' median weights of all measures are given in Table 1 and the level weights for the graded measures (moderate vs severe) are given in Table 2. Expert opinion showed biggest discrepancies within weights for the measures calf mortality and proportion of abattoir remarks for the secondary animal-based measures; the scraping system (exposed vs covered) for resource-based measures; and lameness and rising behaviour for the primary animal based-measures.

Descriptive results show relatively little spread of AWI scores among herds across all information sources (Table 3).

Table 2 Weights assigned by expert opinion to the two levels of moderate or severe impairment of 11 graded welfare measures for an Animal Welfare Index for dairy cattle based on clinical and behavioural measures (modified after Burow *et al* 2013).

Measure	Weight ratio (moderate vs severe)
Lameness	0.33 vs 1
Integument alterations (carpus, tarsus, body)	0.33 vs 1
Body condition	0.33 vs 1
Rising behaviour	0.33 vs 1
Hair coat	0.50 vs 1
Hygiene (hind, udder)	0.50 vs 1
Avoidance distance	0.50 vs 1
Hygiene leg	0.67 vs 1

Associations between AWIs based on three different sources of information

Significant association was found between the AWI 3 and the AWI 1 for data from the period of 180 days prior to visit ($P = 0.04$; R^2adjusted = 0.04) as shown in Table 4.

Register-based key indicators for animal-based measures and AWI 3

Exploring associations between single register-based predictive key indicators and single animal-based measures showed similar patterns as most associations were found between the animal-based measures and milk yield, abattoir and mortality data (Table 5; see supplementary material to papers published in *Animal Welfare* on the UFAW website: http://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). Regarding the prediction of AWI 3, the multivariable analysis showed very limited predictive potential as only the proportion of abattoir remarks remained in the final prediction model for the time periods 365 and 180 days both with a $P < 0.01$ (R^2adjusted = 0.19). The prediction model for the shortest period of time prior to farm visits contained cow mortality, calf mortality, mean annual milk yield in kg energy-corrected milk (ECM) and the standard deviation in ECM for first lactation cows ($P < 0.001$, R^2adjusted = 0.21).

Discussion

The aim of the present study was to investigate the associations between three welfare indices based on different information sources. This was done by applying the same model framework for the three Animal Welfare Indices (AWIs) using an additive weighted model, where weights were derived by expert opinion. For the register-based AWI, three different time-periods for index creation were assessed. Significant negative association was found between the register-based AWI (AWI 1) for the period of 180 days prior to the on-farm evaluation and the animal-based AWI (AWI 3). Despite the relative small variations

doi: 10.7120/09627286.25.2.207

Table 3 Descriptive summary statistics for Animal Welfare Index (AWI) scores for dairy cattle based on different information sources.

Variable	Mean (± SEM)	Median	Minimum	Q1	Maximum	Q3	Possible range
AWI 1 365 days	70.3 (± 16.69)	69.00	41.00	57.00	125.00	78.00	0–196.5
AWI 1 180 days	69.93 (± 14.34)	69.00	41.00	61.00	119.00	76.00	0–196.5
AWI 1 90 days	73.51 (± 12.73)	71.00	49.00	65.00	108.00	81.00	0–196.5
AWI 2	33.23 (± 5.16)	33.50	18.00	30.50	43.50	36.24	0–52
AWI 3	944.65 (± 262.17)	937.85	449.99	741.19	1,757	1,096.02	0–3,900

AWI 1 = register-based measures; AWI 2 = resource-based measures; AWI 3 = animal-based measures and for different time-periods prior to data collection on-farm for resource- and animal-based measures (AWI 1); n = 73 herds.

Table 4 The results of the linear regression models assessing associations between the Animal Welfare Index (AWI) scores for three different information sources (AWI 1 register-based, AWI 2 resource-based and AWI 3 animal-based) for dairy cattle.

	Resource-based AWI 2		Animal-based AWI 3	
	P-value	R^2_{adj}	*P*-value	R^2_{adj}
AWI 1a	0.11	0.02	0.49	–0.007
AWI 1b	0.63	–0.01	0.04*	0.04
AWI 1c	0.12	0.02	0.58	–0.01
AWI 2	–	–	0.26	0.004
AWI 3	–	–	–	–

Significant associations are highlighted by an asterix (* $P < 0.05$) and the adjusted coefficient of determination (R^2_{adj}) stating the degree of variation explained by the given model.

in AWI 1 scores between periods, no significant associations were found for the other two time-periods for AWI 1 and the AWI 3 or between the resource-based AWI 2 and the AWI 3. Although a significant relationship was found, only 4% of the variation within the outcome of the AWI 3 could be explained by the AWI 1, indicating poor model fit and poor predictive ability of the register-based index. The findings of the current study highlight the challenges in finding alternative and cheaper welfare assessment methods, than the laborious yet more direct animal-based assessments. A parallel study performed by Knage-Rasmussen *et al* (2015), pursuing a similar objective for sows, could not find any linear dependency either. The reasons for the lack of associations could be discussed from many angles, eg validity of measures and the aggregation of measures as seen in the following paragraphs. Hence, the present study does not support the idea of assessing animal welfare without visiting the farm.

Selection of measures

In contrast to the WQ® protocol, which presents an integrated approach covering different data sources, such as routine registrations, resource measurements as well as 20 direct animal observations, the present study investigated each data source individually. Additionally, for the purpose of increased feasibility, the number of animal-based measures was heavily reduced to eight measures. Almost all behavioural measures within the WQ® were excluded, namely the Qualitative Behaviour Assessment and social/agonistic behaviour. This was done in order to reduce time, also taking into account results from previous studies on social behaviour in Danish dairy herds showing very low frequencies of agonistic behaviour (Burow *et al* 2009; Andreasen *et al* 2013). Measures for nasal and ocular discharge were discarded, as these have a very low frequency in Danish dairy herds and are very season-dependent measures. Herd visits were conducted across all four seasons and not distributed evenly among seasons; hence, adjustment for seasonality was not possible. Lying down behaviour was initially recorded but due to missing data, the measure was excluded. Getting up behaviour was used instead. The other animal-based measures (hair coat, getting up behaviour, claw conformation) were consistent with the welfare protocol by the Danish Cattle Federation (DCF). A study by Andreasen *et al* (2014) showed significant correlation between the overall WQ® score of farms and an extended DCF protocol. In this extended version, measures of BCS, lying down and collisions with equipment and lameness were assessed as in the WQ®, as was also done in the present study. Furthermore, Andreasen *et al* (2014) included avoidance distance and register data for incidence of dystocia, mortality and milk somatic cell count. Using the WQ® as a gold standard might have shown different patterns of significance; however, the presently used protocol was considered valid for determining animal welfare under Danish production settings.

Danish legislation requires the use of analgesics when dehorning and tail-docking and, furthermore, tail-docking is only allowed upon medical indication, not as a preventive management procedure. Therefore, the measures reflecting the criteria 'absence of pain by management induced procedures' were not applicable under Danish production settings. Hence, not all 12 criteria were covered by the measures included in the present AWI 1, 2 and 3. The register-based measures only covered a part of all criteria, as only criteria

https://doi.org/10.7120/09627286.25.2.207 Published online by Cambridge University Press

'absence of prolonged hunger', 'absence of injuries', and 'absence of disease' were represented. In contrast, the animal-based measures covered the broadest range with the criteria: 'absence of prolonged thirst', 'absence of prolonged hunger', 'comfort around resting', 'absence of injuries', 'absence of disease' and 'good human-animal relationship'. The choice of operating on principle level rather than on criteria levels like the WQ® could, however, have enabled the assessment of more direct associations between the different indices as shown by Andreasen *et al* (2014).

Another discrepancy between the present indices lies in the nature of measures, ie with the AWI 1 being driven by incidence and AWI 3 by prevalence. The cross-sectional results are very time-dependent and even small changes within one measure could alter the AWI 3 substantially if weighted as highly important. This was exemplified by Kirchner *et al* (2014) investigating the WQ® in beef bull farms at three different points in time showing differing values on both measure level (eg mean prevalences), criterion and principle scores. Finally, the wider-ranging inter-observer agreement levels for the different measures might have led to an over- or underestimation of the true prevalence as shown by Otten *et al* (2013), influencing the final herd AWI 3 herd scores. Balancing study designs and ongoing calibrations are inevitable in these large-scale, cross-sectional studies to ensure good data quality and validity of the models.

Aggregation of measures

Major challenges lie in defining, selecting and aggregating welfare indicators based on the existing data sources that cover all aspects of animal welfare of biological functioning, affective state and natural living (Fraser *et al* 1997) and which at the same time are well correlated with the actual welfare status of the given herd. This resulted in the Welfare Quality® development of a comprehensive and multi-dimensional welfare index. A comprehensive index based on a state of the art assessment protocol. However, the lack of transparency became clear as regards to the aggregation of measures into an overall score. Much effort was put into the modelling part, much of this based on expert opinion. As reported by Bonde *et al* (2009), it was found that experts never followed linear curves when asked to score virtual datasets and expert answers were therefore modelled by non-linear utility functions, based on which experts' measure weightings were indirectly defined. But, as seen in retrospect, WQ® researchers concluded that although underlying calculations had been explained in detail these appeared very complex (for example, in Veissier *et al* [2011], and Welfare Quality® [2009a,b,c]) and therefore were only accessible for a very narrow target audience. Hence, in order to obtain transparency, the current study chose a simple linear model also based on expert opinion using direct and indirect estimation techniques for scaling as proposed by Scott *et al* (2001). However, the interpretation of AWI scores based on sums can hold potential hazards, as it does not regard compensation and trade-off between measures or categories (Botreau *et al* 2007) as welfare advantages are not included as negative or deductive variables. However, Botreau *et al* (2007) argued that the use of such sums could be beneficial for assessments of welfare subsets, an approach also used by the Swedish Dairy Association (Sandgren *et al* 2009) using seven different focus areas throughout the dairy cow lifecycle to assess the on-farm welfare.

The current aggregation model has previously been validated by Burow *et al* (2013) investigating overall dairy cattle welfare in herds under grazing and barn conditions. The overall welfare outcomes changed accordingly, when prevalences of the included clinical measures changed from indoor to pasture conditions. However, as mentioned by Burow *et al* (2013), the major drawback of the summation approach is its vulnerability as regards changes in prevalence for the strongly weighted measures. The current AWI will be strongly driven by the highly dynamic measures of lameness, hock and carpal lesions and lean cows and these fluctuations in AWI could potentially alter the pattern of associations with the register-based AWI. Hence, predictions based on single, register-based measures or a subset of measures will be more inconsistent over time than predictions based on a full index. Nonetheless, a single, additive index score has a biologically plausible foundation as single indicators can relate to more than one welfare criteria, eg 'lameness' can be caused by injuries as well as being of infectious origin. As an important feature, the aggregated model for the given indices should also be adaptable to changes in time, eg the expert weights should be revised. Nonetheless, results for expert weights in the present study are similar to a previous study as experts ranked lean animals, mortality rates, wounds and fractures as being of most importance for impaired sow welfare (Knage-Rasmussen *et al* 2015).

A distinct feature in the current study was the moderate spread of both AWI scores and the single measures included in the AWIs across herds. This could be caused by the somewhat biased sampling of herds among positive respondents in the initial survey. The inclusion criteria for the herds were set to be as representative of the Danish dairy cow population as possible. But it is very likely that this has influenced the moderate spread, as the sample did not contain any extreme differences in resource-based measures, as identical production systems were evaluated. However, the AWI 2 did cover the widest range by covering 49% of the possible AWI spectrum compared to the 34% coverage by the AWI 3. Looking for differences in similar herds is a very difficult task and the moderate spread between AWIs in herds is most likely to be held responsible for the lack of association as also documented by Andreasen *et al* (2013). The lack of association between the resource-based and the primary animal-based measures indicate that management has a major impact on dairy cattle welfare. This is illustrated by the discrepancies between scores, as cows in a 'bad' system can have acceptable welfare scores (based on animal-based measures) and *vice versa*. It also emphasises the multi-complexity of animal welfare as the simple provision of resources does not necessarily guarantee good welfare as a resource-based welfare assessment might imply in the given case.

Animal welfare implications and conclusion

In conclusion, the limited associations between animal welfare assessed by a register-based animal welfare index and the direct animal welfare index based on observations of the animals do not support the hypotheses of remote welfare assessment. The time-dependent fluctuations in incidences of register measures do not render this register-based welfare index with high validity, as the register-based index was only significantly associated with the animal-based welfare index for a time-period of 180 days. Further investigations are needed, not only in order to find the most predictive and robust combination of measures and refining the current protocols, but also to determine which time-period the register-based measures should be based on. In this study, three different time-periods were used for extracting register data and further research determining the most predictive time-period is needed. Dairy herds are very volatile entities and specific actions produce different impacts and consequences; hence, a rolling average and aberrations from this might have better perspectives in terms of predicting animal welfare than using means from fixed time-periods. However, the use of cheaper and more feasible welfare measures, ie secondary animal-based register data and resource-based, should only be used for screening purposes, while the final welfare assessment should consist of primary animal-based measures complemented by additional measures (register- and resource-based) to ensure coverage of all aspects of welfare.

References

Andreasen S, Sandøe P and Forkman B 2014 Can animal-based welfare assessment be simplified? A comparison of the Welfare Quality® protocol for dairy cattle and the simpler and less time-consuming protocol developed by the Danish Cattle Federation. *Animal Welfare 23*: 81-94. http://dx.doi.org/10.7120/09627286.23.1.081

Andreasen S, Wemelsfelder F, Sandøe P and Forkman B 2013 The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. *Animal Welfare 143(1)*: 9-17. http://dx.doi.org/10.1016/j.applanim.2012.11.013

Bonde M, Boutreau R, Bracke MBM, Butterworth A, Capdeville J, De Rosa G, Engel B, Keeling LJ, Perny P, Rousing T, Sørensen JT, Van Reenen K and Veissier I 2009 *Final model of multicriterion evaluation of animal welfare.* http://pure.au.dk/portal/files/3054923/2009_06_17_D2_28_deliverable_WP2_3_Final_model_of_multicriterion_evaluation_of_animal_welfare.pdf

Botreau R, Veissier I, Butterworth A, Bracke MBM and Keeling LJ 2007 Definition of criteria for overall assessment of animal welfare. *Animal Welfare 16*: 225-228

Burow E, Rousing T, Halekoh U and Knierim U 2009 Social interactions of dairy cows introduced postpartally to a separated barn section – pilot study. *Acta Agriculturae Scandinavica Section A 59*: 192-196. http://dx.doi.org/10.1080/09064700903160710

Burow E, Rousing T, Thomsen PT, Otten ND and Sørensen JT 2013 Effect of grazing on the cow welfare of dairy herds evaluated by a multidimensional welfare index. *Animal 7(5)*: 834-842. http://dx.doi.org/10.1017/S1751731112002297

Byrt T, Bishop J and Carlin JB 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46*: 423-429. http://dx.doi.org/10.1016/0895-4356(93)90018-V

DeVries M, Bokkers EA, Dijkstra T, van Schaik G and de Boer IJ 2011 Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. *Journal of Dairy Science 94*: 3213-3228. http://dx.doi.org/10.3168/jds.2011-4169

DeVries M, Bokkers EA, van Schaik G, Engel B, Dijkstra T and de Boer IJM 2014 Exploring the value of routinely collected herd data for estimating dairy cattle welfare. *Journal of Dairy Science 97(2)*: 715-730. http://dx.doi.org/10.3168/jds.2013-6585

Dewey C, Haley C, Widowski T, Friendship R, Sunstrum J and Richardson K 2009 Using data collected for production or economic purposes to research production animal welfare: An Epidemiological Approach. *Journal of Applied Animal Welfare Science 12(2)*: 105-113. http://dx.doi.org/10.1080/10888700902719781

European Food Safety Authority (EFSA) 2012 Scientific Opinion of the Panel on Animal Health and Welfare on Guidance and Risk Assessment for Animal Welfare. *EFSA Journal 10(1)*: 2513

Fraser D, Weary DM, Pajor EA and Milligan BN 1997 A scientific conception of animal welfare that reflects ethical concerns. *Animal Welfare 6*: 187-205

Keeling L 2009 *An Overview of the Development of the Welfare Quality Project© Assessment Systems. Welfare Quality Reports no 11*. Cardiff University: UK

Kelly PC, Moore SJ, Blake M and Hanlon AJ 2011 Identification of key performance indicators for on-farm animal welfare incidents: possible tools for early warning and prevention. *Irish Veterinary Journal 64*: 13. http://dx.doi.org/10.1186/2046-0481-64-13

Kirchner MK, Schulze Westerath H, Knierim U, Tessitore E, Cozzi G, Pfeiffer C and Winckler C 2014 Application of the Welfare Quality® assessment system on European beef bull farms. *Animal 8(5)*: 827-835. http://dx.doi.org/10.1017/S1751731114000366

Knage-Rasmussen KM, Rousing T, Sørensen JT and Houe H 2015 Assessing animal welfare in sow herds using data on meat inspection, medication and mortality. *Animal 9(3)*: 509-515. http://dx.doi.org/10.1017/S1751731114002705

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

Mulleder C, Troxler J, Laaha G and Waiblinger S 2007 Can environmental variables replace some animal-based parameters in welfare assessment of dairy cows? *Animal Welfare 16(2)*: 153-156

Nyman AK, Lindberg A and Sandgren CH 2011 Can pre-collected register data be used to identify dairy herds with good cattle welfare? *Acta Agriculturae Scandinavica, Section A – Animal Science 53 (S1)*: S8

https://doi.org/10.7120/09627286.25.2.207 Published online by Cambridge University Press

Otten ND, Nielsen LR, Thomsen PT and Houe H 2014 Register-based predictors of violations of animal welfare legislation in dairy herds. *Animal 8(12)*: 1963-1970. http://dx.doi.org/10.1017/S1751731114001918

Otten ND, Toft N, Houe H and Thomsen PT 2013 Adjusting for multiple clinical observers in an unbalanced study design using latent class models of true within-herd lameness prevalence in Danish dairy herds. *Preventive Veterinary Medicine 112(3-4)*: 348-354. http://dx.doi.org/10.1016/j.prevetmed.2013.09.002

R Development Core Team 2012 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. http://www.R-project.org

Rousing T, Jakobsen IA, Hindhede J, Klaas IC, Bonde M and Sørensen JT 2007 Evaluation of a welfare indicator protocol for assessing animal welfare in AMS herds: Researcher, production advisor and veterinary practitioner opinion. *Animal Welfare 16*: 213-216

Sandgren CH, Lindberg A and Keeling LJ 2009 Using a national dairy database to identify herds with poor welfare. *Animal Welfare 18*: 523-532

Scott EM, Nolan AM and Fitzpatrick JL 2001 Conceptual and methodological issues related to welfare assessment: A framework for measurement. *Acta Agriculturae Scandinavica Section A- Animal Science 30*: 5-10. http://dx.doi.org/10.1080/090647001316922983

Thomsen PT, Østergård S, Sørensen JT and Houe H 2007 Loser cows in Danish dairy herds: Definition, prevalence and consequences. *Preventive Veterinary Medicine 79*: 116-135. http://dx.doi.org/10.1016/j.prevetmed.2006.11.011

Veissier I, Jensen KK, Botreau R and Sandøe P 2011 Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. *Animal Welfare 20(S1)*: 89-101

Welfare Quality® 2009a *Welfare quality assessment protocol for cattle*. Welfare Quality® Consortium: Lelystad, Netherlands

Welfare Quality® 2009b *Welfare quality assessment protocol for pigs (sows and piglets, growing and finishing pigs)*. Welfare Quality® Consortium: Lelystad, Netherlands

Welfare Quality® 2009c *Welfare quality assessment protocol for poultry*. Welfare Quality® Consortium: Lelystad, Netherlands

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