Preventive Veterinary Medicine 213 (2023) 105880

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Automated monitoring and detection of disease using a generic facial feature scoring system – A case study on FMD infected cows

Gerben Hofstra[a], Hilde van Abeelen[a], Marleen Duindam[a], Bas Houben[a], Joris Kuijpers[a], Tim Arendsen[b], Mathijs van der Kolk[b], Felix Rapp[b], Jessy van Spaendonk[b], José L. Gonzales[c], Ronald Petie[c,*]

[a] HAS University of Applied Science, Onderwijsboulevard 221, 5223 DE 's-Hertogenbosch, the Netherlands

[b] AVANS University of Applied Science, Onderwijsboulevard 215, 5223 DE 's-Hertogenbosch, the Netherlands

[c] Epidemiology Bioinformatics and Animal Models, Wageningen Bioveterinary Research, Houtribweg 39, 8221 RA Lelystad, the Netherlands

ARTICLE INFO

Keywords:
Cattle, face, foot-and-mouth disease
Images
Image recognition
Disease monitoring
Disease detection

ABSTRACT

Digital images are becoming more readily available and possibilities for image processing are developing rapidly. This opens the possibility to use digital images to monitor and detect diseases in animals. In this paper we present 1) a generic facial feature scoring system based on seven facial features, 2) manual scores of images of Holstein Frisian heifers during foot-and-mouth disease vaccine efficacy trials and 3) automatic disease scores of the same animals. The automatic scoring system was based on the manual version and trained on annotated images from the manual scoring system. For both systems we found an increase in disease scores three days post infection, followed by a recovery. This temporal pattern matched with observations made by animal caretakers. Importantly, the automatic system was able to discern animals that were protected by the vaccine, and did not develop blisters at the feet, and animals that were not protected. Finally, automatic scores could be used to detect healthy and sick animals with a sensitivity and specificity of 0.94 on the second and third days following infection in an experimental setting. This generic facial feature disease scoring system could be further developed and extended to lactating Holstein Frisian dairy cows, other breeds and other infectious diseases. The system could be applied during animal experiments or, after further development, in a farm setting.

1. Introduction

The aim of Precision Livestock Farming (PLF) is to optimize the management of animals by continuous real-time monitoring of health, welfare, production, and environmental impact (Berckmans, 2017). Many PLF technologies such as accelerometers, automated milking robots and position sensors can continuously monitor each individual, recording behaviour, productivity and current health status. In addition, the algorithms within these systems can learn at the individual level, learning what is normal for each animal (Berckmans, 2017) and signal deviations in case of disease (Mandel et al., 2017; Stangaferro et al., 2016a, 2016b). Many of the goals of PLF are also applicable in a research setting, adding continuous monitoring of sensor data from animals to the research tools.

One potential tool is the use of automated scoring of facial expressions. Facial expressions in animals vary widely across species, but there are species-specific patterns that can be used to explore the emotional state of the animal (Descovich et al., 2017). It is recognized by researchers that emotions are part of the dairy cow life (Proctor and Carder, 2015). Research has for instance shown that eye white and ear posture are potential promising indicators for interpreting emotions in ruminants. (Battini et al., 2019; Lambert (Proctor) and Carder, 2017; Proctor and Carder, 2015, 2014; Reefmann et al., 2009; Sandem and Braastad, 2005).

Foot-and-mouth disease (caused by virus family Picornaviridae) is an infectious disease in cloven-hoofed animals which could induce mild to acute disease in infected cattle. Characteristic clinical signs in diseased animals are fever and vesicular lesions of the feet, tongue, snout and

Abbreviations: FMD, Foot-and-mouth disease; PLF, Precision Livestock Farming; IBR, Infectious bovine rhinotracheitis; IBK, Infectious bovine keratoconjunctivitis; DPI, Day post infection; LMM, Linear regression mixed effects model; RMSE, Root Mean Squared Error.

* Corresponding author.
E-mail address: ronald.petie@wur.nl (R. Petie).

https://doi.org/10.1016/j.prevetmed.2023.105880
Received 8 June 2022; Received in revised form 15 February 2023; Accepted 17 February 2023
Available online 18 February 2023

teats. Wageningen Bioveterinary Research performs regular foot-and-mouth disease (FMD) vaccine efficacy trials for vaccine developers. The availability of the vaccine efficacy trials and the nature of the symptoms affecting the face of the animal makes FMD a suitable disease for a proof of concept of a generic facial feature scoring system.

In this paper we explore the question whether it is possible to monitor and detect disease using a generic facial feature scoring system. We applied the scoring system to five FMD trials. A manual scoring system was developed, based on a literature review and tested and validated on a set of heifer face photos taken during a foot-and-mouth disease vaccine efficacy trial. An automatic system was developed to automatically score disease based on images of heifer faces using image recognition software.

2. Material and methods

2.1. Literature search

To select the features for the face-based scoring system, a literature search was performed combining search terms part 1 and 2 from Table 1. The Constrained Snowball Sampling method was used to filter the results by relevance (Lecy and Beatty, 2012) and the selected articles formed the conceptual basis of the scoring system.

2.2. Generic facial feature scoring system

Generic, face-based disease indicators can be based upon features of the eyes, ears, mouth and nose. Disease indicators were general indicators, identified independently of whether they can be applied to an FMD infection. The literature research resulted in indicators for ocular discharge, corneal opacity, eye position, ear position, nasal discharge, drooling and wounds and other abnormalities of the skin of the head. A score of 0 was assigned to positive and/or neutral aspects, while a score of 1 was given for intermediate expression and 2 for full expression of negative aspects. Scores were then weighed by multipliers, which strengths were based on the potential impact of associated animal diseases on the individual, herd or population. With the aim of making the system more sensitive to contagious diseases. The scoring system was developed using the images from the first experiment. Images from the second experiment were used as a manual implementation case of the scoring system (Table 2).

2.3. Eyes

Three disease indicators were based on the eyes: ocular discharge, corneal opacity and eye position. Presence of ocular discharge can result from irritation to the cornea or conjunctiva, or both, and could indicate illnesses in cattle, such as infectious bovine rhinotracheitis (IBR) and infectious bovine keratoconjunctivitis (IBK). These are contagious diseases which not only affect the individual animal but can also have effects at population level (Brown et al., 1998; Yates, 1982); for these reasons ocular discharge received a multiplier of 2. To score ocular discharge, three scores were used: 0, no ocular discharge present; 1, small to moderate amount of ocular discharge below at least one of the eyes and 2, heavy ocular discharge below and above at least one of the eyes (McGuirk and Peek, 2014).

Besides ocular discharge, corneal opacity can also be an indicator of disease. As mentioned above, diseases such as IBK or IBR can cause the eye to get infected. When the eye is infected, the cornea usually starts to get cloudy within the first 24 h after infection. Without treatment, the eye will turn white/yellow due to the presence of puss and could eventually become blind. IBK is a highly contagious disease, causing inflammation of the cornea (the clear outer layer) and conjunctiva (the pink membrane lining the eyelids) of the eye (Angelos, 2020). Because IBK and IBR are highly contagious (Brown et al., 1998; Yates, 1982) and corneal opacity is a more severe symptom than ocular discharge, corneal opacity got a multiplier of 3. In addition, a study of Starič, Križanec, and Zadnik ((2008) showed no correlation between ocular discharge and corneal opacity when cows were infected with listeria monocytogenes keratoconjunctivitis and uveitis. Therefore, it is useful to include corneal opacity in the scoring system, in addition to ocular discharge, because some diseases are missed when only ocular discharge is considered. Two scores were used: 0, dark, transparent eyes and 2, bluish to white opacity in at least one of the eyes.

The position of the eyes with respect to the eye socket can indicate dehydration or extreme emaciation (Peek and Divers, 2018). If several cows suffer from dehydration or extreme emaciation, this can be a structural problem on the farm. Therefore, this indicator was given a multiplier of 2. Two scores were used: 0, a normal eye bulge; 2, abnormal eye bulge in any severity (> 5 mm deviation in eye bulge).

Table 1
Search terms used in literature search.

Search term part 1	Search term part 2
Stress symptoms	Cow
Facial expressions	Cattle
Scoring method welfare	Ruminant
FMD symptoms	Animal
Facial indicators welfare	Livestock
Subjective scoring	Calves
Five freedoms	Heifer
Emotions	Dry cows
Eye white	Dairy cows
Ear posture	Bovine
Pain evaluation	Herd
Subjective scoring	Horse
Health and welfare	Goat
Health scoring	Sheep
Illness symptoms	Pig
Valence and arousal	Rat
Body language	
Behaviour assessment	

2.4. Ears

There are indications that ear posture may be useful in assessing emotional valence in animals (Proctor and Carder, 2014; Reefmann et al., 2009), where valence is the positive or negative quality of the experience (Mendl et al., 2010). In ruminants positive emotions are behaviourally expressed by high proportions of non-erect ear posture, such as horizontal, backward-down ear posture and right-ear-asymmetry ear posture (Goma et al., 2018; Tamioso et al., 2017). Cows in a neutral emotional state are associated with the horizontal, or hanging, ear posture (Proctor and Carder, 2014). Long durations of hanging ears were observed when the animals experienced social licking (Schmied et al., 2008). This implies that, hanging ears are associated with positive emotional states and low arousal (Battini et al., 2019). In addition, the backward-down ear posture is linked with positive emotional states and low arousal in farm animals (Battini et al., 2019; Lambert and Carder, 2019; Tamioso et al., 2017). Asymmetry of ear posture may differ with negative and positive emotional states. For example, right-ear-asymmetry ear posture is common during brushing and feeding of dairy cows (positive emotional state). Left-ear-asymmetry ear posture is common during sudden and stressful situations with an associated negative emotional state (Boissy et al., 2011; de Oliveira and Keeling, 2018) and is associated with activation of the right hemisphere of the brain (de Oliveira and Keeling, 2018).

In ruminants, negative emotions are behaviourally expressed by high proportions of raised-up ear posture, forward-ear posture, backward-up ear posture and left-ear-asymmetry ear posture. The raised-up ear posture and forward-ear posture represent negative valence with a high arousal (Boissy et al., 2011; Tamioso et al., 2017). The backward-up ear posture is associated with unfamiliar and uncontrollable unpleasant situations, such as pain (Gleerup et al., 2015). A severely sick cow

Table 2
Overview of the setup of the FMD vaccine efficacy trials and the annotated images available.

Experiment	Purpose	Number of animals	Protected (§)	Number of images	Treatment	Animals
1	Develop scoring system. Train neural network.	22	14	189	Controls	2
					Vaccinated	20
2	Assess manual scoring system. Train neural network.	17	10	236	Control	2
					Vaccinated	15
3	Train neural network.	22	12	169	Control	2
					Vaccinated	20
4	Train neural network.	17	11	155	Control	2
					Vaccinated	15
5	Unseen images used for validation	18	18	18	Control	0
					Vaccinated	18

§ Animals were considered protected by the vaccine if blisters were absent from the claws.

generally has ears that droop down below horizontal, which can be caused by depression, pain, or fever (Smith and Risco, 2005). From ear posture alone it cannot be determined whether the cause is contagious and therefore a risk for the individual or the entire population. Therefore, ear posture was given a multiplier of 2. Three scores were used: 0, positive or neutral ear posture score; 1, negative ear posture and 2, ear postures associated with sickness. Positive or neutral ear postures are ears backward and down, ears axial, left ear axial and right ear backward. Ears held upright above the head and neck, ears forward or right ear axial and left ear backward are considered negative ear postures. Ears drooping below horizontal are associated with sickness.

2.5. Muzzle

When bacteria and viruses infect cows through the respiratory tract, nasal discharge can occur. Nasal discharge can also be the result of the accumulation of normal respiratory secretions when the muzzle is not cleaned by the animal due to disease. Excessive or abnormal discharge is usually an indication of health problems (Donaldson et al., 1970; McGuirk and Peek, 2014; Rice et al., 2008). Because respiratory diseases tend to be contagious and FMD for instance is a notifiable disease, the severity of nasal discharge is high and was given a multiplier of 3.

Different scoring methods can be used for nasal discharge. The Welfare Quality Assessment Protocol (Welfare Quality, 2009) for cattle uses a scoring method where nasal discharge is classified in two categories: ''no evidence of nasal discharge'' or ''evidence of nasal discharge''. In this Welfare Quality Assessment protocol, nasal discharge is defined as clearly visible flow/discharge from the nostrils: transparent to yellow/green and often of thick consistency. In a study by McGuirk and Peek (2014) a scoring method is used with four categories to diagnose respiratory disease in calves. As the muzzle of a healthy animal is moist, it is important that moist is not confused with the presence of a small amount nasal discharge. In case it does happen, it is important that the confusion does not immediately affect the total score of the cow. Therefore, three values were used in this study: score 0, normal serous discharge in at least one nostril; score 1, small to intermediate mucus discharge in at least one nostril and score 2, copious mucopurulent discharge in at least one nostril.

Drooling is a symptom of FMD (Alexandersen et al., 2003), as well as bluetongue (Goltz, 1978) and severe heat stress (Kadzere et al., 2002). Because FMD and bluetongue are OIE listed diseases, the severity of drooling is high and is given the high multiplier of 3. The absence of drooling is scored as 0 and presence of drooling is scored as 2. To avoid confusing nasal discharge with drooling, drooling is defined as foam or wet appearance around the lips.

2.6. Skin abnormalities

Wounds and other abnormalities of the skin are included in this scoring system. Especially when they are the result of an infection like bovine sarcoptic and psoroptic mange, which is an infection that spreads easily to other cows in the herd (Hamel et al., 2015). Cows are motivated to get rid of the itch by scratching when suffering from diseases such as mange. They can do this by rubbing against rough surfaces, which can cause wounds (Moncada et al., 2020). Usually facial wounds are temporary, and animals heal quickly without experiencing too much discomfort. When cows are infected with FMD or Vesicular Stomatitis Virus (VSV), blisters can be visible in and around the mouth (de Oliveira et al., 2018). However, since FMD is but one of many causes of skin abnormalities, this trait was given a multiplier of 1. To score wounds and other abnormalities of the skin, three values were used: 0, small lesions or hairless patches or no hair loss or lesions; score 1, one or more large hairless patches and score 2, one or more large lesions (areas of skin damage, wound or scab) and/or scratches or cuts. Patches and lesion are considered large when they measure approximately 2 cm in diameter or more.

2.7. Foot-and-mouth disease trials

Foot-and-mouth disease (FMD) vaccine efficacy trials are performed under veterinary Bio Safety Level 4 (v-BSL4) conditions at Wageningen Bioveterinary Research (WBVR) in Lelystad, the Netherlands. FMD vaccine trials are routinely performed at [institute acronym], following strict ethical procedures and approval.

For all experiments the heifers aged approximately 9 months were introduced into the facilities 25 days before they were challenged by inoculation of FMD virus into the tongue to induce infection. The animals were tethered to a band running from the ceiling to the floor. In the experiments, there were groups of heifers that were not vaccinated before challenge (“controls”) and groups of heifers that received a vaccination at 21 days before challenge. Different vaccination doses were used. Expressions of clinical signs in infected heifers ranged from acute (mostly controls) to mild. Some animals were removed from the trials, because they reached the vaccine trial planned endpoint. The different experiments and how their corresponding images were used in this study are summarised in Table 2.

2.8. Images from foot-and-mouth disease trials

For our study we opportunistically collected images during five FMD trials. As a result, no additional animal experiments were needed for the development of our facial feature scoring system, optimising the use of experimental animals for additional purposes to those primarily intended (vaccine efficacy trials). Note that the images belong to the commercial partners that ordered the FMD trials and cannot be shared by us. At least two images were taken per animal per day. Photos were taken trying to disturb the animals the least as possible. Making the photographs did not require invasive handling of the animals.

2.9. Clinical inspections

During the FMD trials the animals were inspected daily for welfare and expression of clinical signs by the animal caretakers using a protocol. Following inspection, caretakers recorded all clinical observations for each heifer. In the protocol for each animal presence was scored for: pain when eating, nasal discharge, reduction in feed uptake, drooling. Three animal caretakers inspected the animals during the trials. The outcome of the clinical inspections for experiment 2 are shown in Fig. 1. Animals were checked for the presence of blisters at the claws by a veterinarian at 3 or 4 days post infection and at the end of the trial.

2.10. Manual scores

In the manuscript, the manual implementation of the generic facial feature scoring system will be referred to as the manual scoring system. Images from the first FMD vaccine efficacy trial were used to validate the generic, literature-based facial feature scoring system. The images from the second FMD vaccine efficacy trial were used to provide a manual implementation case of our scoring system (Table 2), test the inter-rater reliability and validate the results against observations by the animal caretakers. The images were scored by four people, who were not aware of the FMD infection status of the animal on the image. One of the observers had missing observations and was omitted from the analysis, resulting in three observers.

The final score for each image was calculated by multiplying the score for each indicator with the corresponding multiplier and taking the sum of the resulting products. The multipliers are outlined above and summarised in Table 3.

2.11. Automatic scores

In the manuscript, the automated implementation of the generic facial feature scoring system will be referred to as the automatic scoring system. For the automatic analysis, images from the third and fourth FMD trial were scored to increase the number of annotated images (Table 2). The annotation was done by a fifth rater. Deep learning networks were trained to score four out of the seven indicators: ocular discharge, nasal discharge, drooling and ear posture. Wounds, corneal opacity and eye position could not be scored automatically because they occurred too infrequent. In total 749 annotated images were available for training the networks.

Annotated images of heifer faces were provided as training material for the neural network, including the relevant facial features: ears, eyes nose and muzzle. The models were developed using standard object detection and classification models from Microsoft Azure's Custom Vision, which were not customised, and were trained on annotated images from experiment 1–4. Data augmentation was used to prevent overfitting to the limited dataset and images were rotated, mirrored and pixel intensities were changed. The model would first detect the position of the head of the heifer and second detect the position of the facial features mentioned above. Thirdly, the facial features would receive their scores for the different indicators. The eyes would, for example, be used to score three facial indicators: ocular discharge, eye position and corneal opacity.

For the score of ear posture, it was important to discern right and left ears. For this purpose, ears were detected first, and their spatial position was used to distinguish left and right. Ocular discharge was scored from the eye-region of the images. As in the manual system, ocular discharge was scored as present when found in at least one of the eyes. Nasal discharge was scored on images of the muzzle of the heifer, drooling was scored on the area below the muzzle.

The performance of the automatic scoring was assessed by comparing the manual and automatic scores of both seen and previously unseen images. Seen images have been used for training the model and unseen images have not. For the performance on seen images, mean manual scores of 60 training images from the second FMD trial were compared to the automatic scores. For the previously unseen images, scores of 18 images taken during a fifth FMD vaccine efficacy trial were also compared with the automatic scores. The manual scores for the images from the fifth trial were provided by a sixth rater.

Table 3
Summary table with scores and multipliers for each of the facial indicators.

Indicator	Scores	Multiplier
Ocular discharge	0 = no ocular discharge 1 = small to moderate amount of ocular discharge below at least one of the eyes 2 = heavy ocular discharge below and above at least one of the eyes	2
Corneal opacity	0 = dark, transparent eyes 2 = bluish to white opacity in at least one of the eyes	3
Eye position	0 = a normal eye bulge 2 = abnormal eye bulge in any severity (> 5 mm deviation in eye bulge)	2
Ear posture	0 = positive or neutral ear posture score1 = negative ear posture 2 = ear postures associated with sickness	2
Nasal discharge	0 = normal serous discharge in at least one nostril 1 = small to intermediate mucus discharge in at least one nostril 2 = copious mucopurulent discharge in at least one nostril.	3
Drooling	0 = absence of drooling 2 = presence of drooling	3
Wounds	0 = small lesions or hairless patches, or no hair loss or lesions 1 = one or more large hairless patches 2 = one or more large lesions (areas of skin damage, wound or scab) and/or scratches or cuts.	1

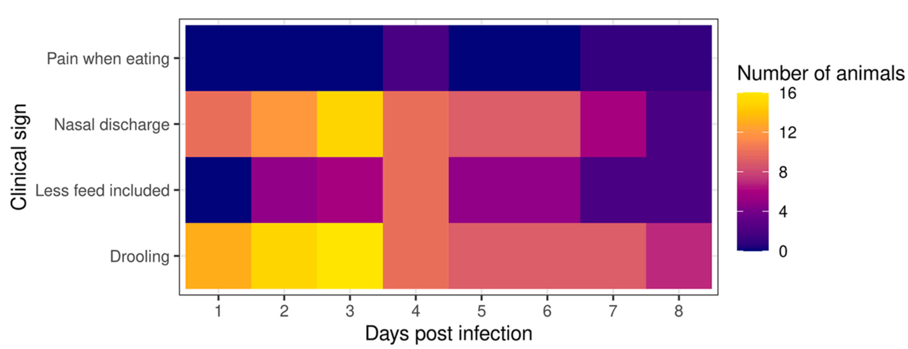

Fig. 1. Temporal trends during clinical inspections of 17 heifers during a foot-and-mouth disease vaccine efficacy trial. Nasal discharge and drooling peaked at three days post infection and the reduction in feed was highest on day four. Note that one animal was removed from the experiment on day three and six animals were removed on day four.

2.12. Statistical analysis

2.12.1. Clinical inspections

The observations of the animal caretakers were presented as the number of animals that showed symptoms per day and were compared to the manual face scores using the Pearson correlation between pairwise complete observations. Complete day pairs were available for day two, three and four post infection.

2.12.2. Inter-rater reliability

For the entire process, from the development of the scoring system to the test of the automatic system, five raters were involved. The inter-rater reliability was assessed twice, once during the testing of the manual scoring system, where the comparison was between the three raters involved (Supplementary table 1). The second inter-rater assessment was between all raters and included the additional two raters that scored images for the automatic scoring system. Fifty randomly chosen images from the second experiment were used.

The inter-rater reliability was tested using Kendall's coefficient of concordance W because the scores were ordinal. Kendall's W was corrected for ties as raters did not rank the indicators. For Kendall's W, a value of 0 indicates no agreement between the raters, while 1 indicates full agreement between the raters.

2.12.3. Manual and automatic face scores

When comparing the manual and automatic scores per facial indicator (Fig. 2), the automatic scores were used directly, but the manual scores were summarised by taking the mean of the scores from the raters. Agreement between the manual and automatic scores was calculated by means of a Pearson correlation test using pairwise complete observations.

2.12.4. Disease dynamics

To assess the ability of the scoring system, either manual or automatic, to reflect the progression of disease severity in time following infection (inoculation) a linear regression mixed effects model (LMM) was used. In this model, the final score was the response variable and the scoring system (manual or automatic), the day post infection (DPI, treated as a categorical variable), and the animal protection status were fitted as explanatory variables. The animals were considered protected by the vaccine if blisters were absent from the claws. Animal ID was fitted as random effect to accounted for repeated measures on the same animals. In the analysis, interactions between system and DPI were also assessed and the protection status was introduced as adjustor variable. Variable significance was assessed using the likelihood ratio test. The threshold for significance was $p < 0.05$. Model fit was assessed by calculating the model's adjusted R^2.

2.12.5. Protection status

The ability of the facial feature scoring system to discriminate disease severity between vaccine protected and not protected animals was tested. This analysis was done using only data of days 2 and 3 post infection. Day 3 prior to infection was excluded because animals were not yet infected, and day 4 post infection was excluded because all the not-protected animals were already removed by this day.

For analysis, a LMM was fitted, where the final scores were the response variable, system, DPI and protection status were the explanatory variables and animal ID was introduced as random effect, like before. Interactions between system and protected and system and DPI were also assessed.

2.12.6. Detection

To test the performance of the score for disease detection, scores from healthy animals at three days before infection were compared to scores of animals at three days post infection. At three days post infection all animals, even the protected ones, display some symptoms, like drooling, either due to disease or the inoculation in the tongue. Different thresholds were tested until the area under the receiver operating characteristics curve was maximized.

All statistical analyses were done using the software package R, version 4.0.3 (R Core Team, 2020) using tidyverse for data manipulation (Wickham et al., 2019). The inter-rater agreement was calculated using the package irr (Gamer et al., 2019), lme4 was used to fit the LMM (Bates et al., 2015) and ROCR was used for the receiver operator characteristics calculations (Sing et al., 2005).

3. Results

3.1. Clinical inspections and manual face scores

The indicators ocular discharge, ear posture, nasal discharge and drooling were given a score other than zero most often by the observers. Wounds, eye position and corneal opacity were observed much less frequently (Table 4).

The clinical inspections by the animal caretakers done on experiment 2 showed that drooling and nasal discharge peaked three days post infection (Fig. 1). In addition, there was a reduction in feed intake. Note that one animal was removed from the experiment on day three and six animals were removed on day four.

Two of the features that were used during the manual facial scoring were scored during the clinical inspections of the animal caretakers too: drooling and nasal discharge. For drooling, the number of animals that

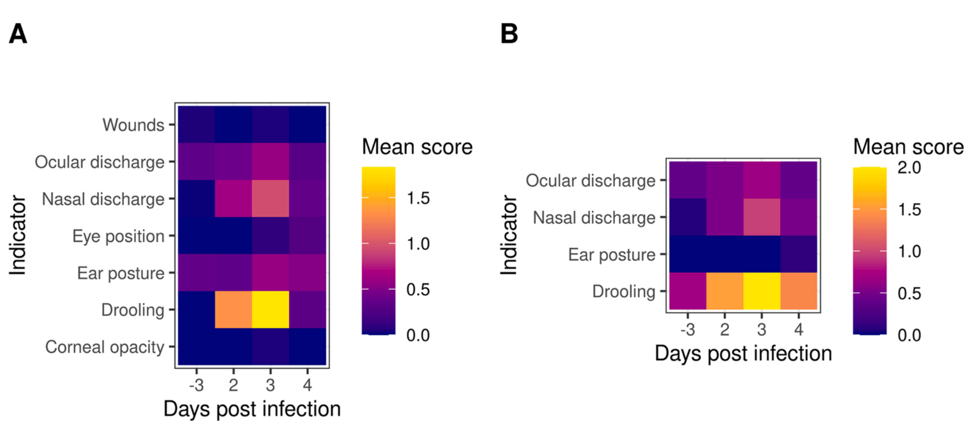

Fig. 2. Comparison of temporal trends in mean scores for in facial disease indicators scored on images of heifer faces. Scoring was done (A) manually by three raters and (B) automatically. Note that only four of the seven indicators could be scored automatically. In general, the trends in the automatic scores match the trend in the manually scored images, with indicators peaking at three days post infection. Drooling was scored more automatically than manually; ear posture was scored less.

Table 4
Frequency of assigning a score other than zero to a facial indicator by any of the observers. The frequency is in number of images.

Indicator	Experiment 1	2	3	4	Total
Ocular discharge	103	102	117	85	407
Ear posture	144	74	97	74	389
Nasal discharge	91	122	84	52	349
Drooling	99	113	66	53	331
Wounds	8	2	17	2	29
Eye position	6	9	8	3	26
Corneal opacity	0	2	7	5	14

exhibited drooling was correlated with the mean manual score with a strength of 0.98. For nasal discharge the correlation was 1.0.

3.2. Inter-rater reliability for all raters

Fifty images of the second FMD trial were scored by all five raters (Table 5). The indicators drooling, nasal discharge and ocular discharge had good W-scores, above 0.6. Moderate agreement existed for ear posture and the agreement for eye position was bad. In the set of images used, corneal opacity and wounds were only scored zero, resulting in 100 % agreement.

3.3. Manual and automatic disease scores

In our facial feature scoring system, the mean scores for the facial features nasal discharge and drooling increased strongly over time and peaked on the third day post infection, for both the manual and automatic scoring systems (Fig. 2). This reflects the clinical observations done by the animal caretakers. The indicators ocular discharge and ear posture showed a similar but weaker trend. The indicators wounds, eye position and corneal opacity displayed only slight variation over time in comparison to the other indicators as they were observed very seldom (see Tables 4 and 5). For this reason, the last three indicators were not included in the automatic score.

The scores for the individual facial indicators were combined into a weighted final score, which was found to peak on day three post infection for both the manual and the automatic scores (Fig. 3). The LMM analysis showed no significant differences between manual and automatic scoring systems ($p = 0.098$) at any of the days post infection. The day post infection had a significant influence ($p < 0.001$) on the expected manual and automatic scores. Expected mean scores increased up to day three post infection, compared to day three before infection, and were followed by a decrease at day four post infection. In addition, an interaction between system and day post infection ($p = 0.033$) was found, where differences in mean scores between the two systems were larger for day –3 and 4 post infection, compared to day 2 and 3.

The mean of both manual and automatic scores were different between heifers considered protected and those that were not ($p = 0.032$). The LMM analysis confirmed no significant differences between manual and automatic scores ($p = 0.26$). No significant interactions between protection status and system ($p = 0.94$), or system and DPI ($p = 0.50$) were observed.

The difference between the predicted automatic scores and the mean manual scores (Table 6) of previously seen images shows that the predictions differ most for the final score, drooling and ear posture. Nasal discharge was predicted with the lowest error in the manual dataset but had a higher error in the automatic system. Performance was lower on unseen images with higher Root Mean Squared Error (RMSE) values.

Table 5
Inter-rater reliability for all five raters, using Kendall's coefficient of concordance W. The Non-zero score indicates the number of images where at least one of the observers assigned a score other than zero.

Indicator	Images	Non-zero	Agreement (%)	W	P_value
Drooling	50	18	92	0.93	$p < 0.001$
Nasal discharge	50	33	50	0.79	$p < 0.001$
Ocular discharge	50	23	56	0.62	$p < 0.001$
Ear posture	50	32	42	0.52	$p < 0.001$
Eye position	50	3	94	0.35	$p < 0.001$
Corneal opacity	50	0	100	–	–
Wounds	50	0	100	–	–

3.4. Detection

The range of final scores is between 0 and 11 for automatic scores. The optimal cutoff value for using the final score for disease detection was 9. Using this threshold resulted in 1 false-positive for healthy animals at 3 days prior to infection and 16 true-negatives. In addition, 1 false-negative and 15 true-positives for were found at 3 days post infection. This translates into a sensitivity and specificity of 0.94 and an area under the receiver operating characteristic curve of 0.96.

4. Discussion

The aim of this work was to develop a generic, face based, disease scoring system for images of cows. This study provided a case study, using solely heifer face images, for which the automatic implementation had a 94 % sensitivity and specificity to detect FMD in a controlled experiment and could distinguish between day two, three and four post infection in a FMD vaccine efficacy trial. The observed temporal pattern in the scores (Fig. 2) corresponded well with the observations of clinical signs made by the animal caretakers (Fig. 1). No significant differences between manual and automatic scoring systems were found and the mean of both manual and automatic scores were different between heifers considered protected and those that were not. This is a promising technique which is worth exploring further as it provides opportunities for automatic scoring of disease based solely on images of cow faces.

In order to extend the validity of the automatic scoring system a couple of approaches should be considered. The current results are based on a group of FMD infected Holstein Frisian heifers and should be extended to lactating Holstein Frisian dairy cows, other breeds and importantly include more diseases. In addition, the low prevalence of the indicators wounds, corneal opacity and eye position in the current dataset was most likely the cause of the low, or missing, coefficient of concordance. The low prevalence of these indicators also prevented the incorporation into the automatic scoring system. However, for the indicators nasal discharge, drooling and ocular discharge a moderate to good rater agreement was found (Table 5).

The overestimation of the indicator drooling within the automatic scoring system (Table 6) could possibly be due to the way the heifers were secured inside the facilities during the trials. The heifers were tethered to a light-coloured vertical strap which, depending on the angle of the image, might be scored as drooling by the automatic scoring system. For future trials a different way of tethering or a more distinguishable colour of the vertical line should be considered. Another explanation for the overestimation of the indicator drooling could lie in the interaction between DPI and system, as observed in our LMM, where the automatic scoring system showed signs of shrinkage towards the mean score resulting in higher automatic scores for days 3 prior to infection and day 4 post infection (Fig. 3).

It should be noted that the technique of inoculation, injection in the tongue, will have caused discomfort of the tongue making it more difficult for all animals to clean their muzzle and swallow saliva, including the animals which were protected by the vaccine. However, the scoring system was still able to distinguish the protected from the non-protected animals. In a natural infection the expression of the different indicators could vary in strength or timing.

Indicator scores were combined into a final score making use of multipliers. Note that the multipliers in the generic facial feature scoring

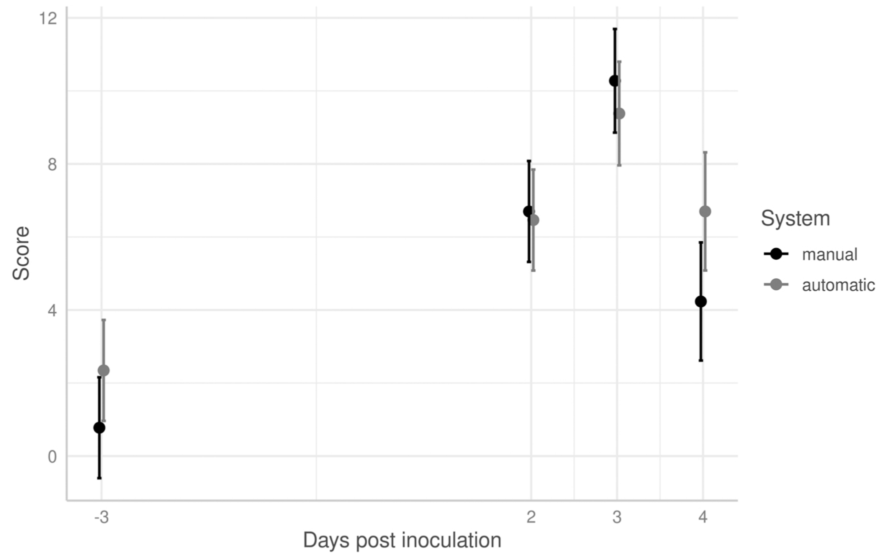

Fig. 3. Temporal trends in final disease scores for the manual and automatic system. Both the manual and automatic final scores peaked at three days post infection and follow the same trend. The difference between systems is largest at day –3 and 4.

Table 6
Difference between manual scores and the automatic score of images used for training, and unseen images. RMSE, Root Mean Square Error.

	RMSE	
Indicator	Seen images	Unseen images
Final score	3.1	3.6
Drooling	1.00	1.3
Ear posture	0.59	0.85
Ocular discharge	0.51	0.58
Nasal discharge	0.35	1.0

system were chosen to reflect the implication that disease could have at a farm, or population level and that the weight of the multipliers affected the amplitude of the final score. Automatic tuning of the multipliers could have increased the performance of both the manual and automatic system, but was purposefully omitted, in order to keep the results interpretable.

More indicators could be added to the scoring system in the future. Visible eye-white for example can be a valuable indicator to gain insight in the emotional state of cows. Cows which are exposed to a positive valence event of low arousal, such as gentle stroking, show a significant decrease in visible eye-white (Proctor and Carder, 2015). Visible eye-white increases when a cow is exposed to a high arousal event, this is, however, unrelated to the type of valence (Lambert (Proctor) and Carder, 2017; Sandem et al., 2004, 2002; Sandem and Braastad, 2005). It is therefore not possible to attribute an increase of visible eye-white to either a negative or positive valence, but this might be possible when context is included (Lambert (Proctor) and Carder, 2017).

Another indicator that might be added to the scoring system in the future is the 'pain face', which is expressed by cows that experience pain (Gleerup, 2017). They display a tense stare, there is tension in the muscles above the eyes and on the side of the head (Gleerup et al., 2015). When cows experience a painful stimulus like dehorning or branding, they show backward ears, dilated nostril, open mouth, raised inner brow, and raised outer brow (in beef cattle) (Müller et al., 2019). In horses, a scale for facial assessment of pain has already been developed and validated (van Dierendonck and van Loon, 2016; van Loon and Van Dierendonck, 2015). These pain indicators were for now deemed too challenging to score by the observers, however the automatic system might be more adept in recognizing this indicator when trained on many images.

Based on the developed system, at least images of the front and the left side of the head are needed for scoring. The left side of the face is important due to the significance of the left ear as an indicator of negative emotional state. It is preferred to also include an image of the right side of the face as to not miss indicators such as wounds, ocular discharge and corneal opacity on the right side of the head. Lighting conditions should also be standardized, preferably using multiple light sources to prevent the casting of shadows.

A future technical improvement could be to use video recordings of cow faces as a basis for the disease score. This would enable the system to use cursory ear movement, seeing that ear movement frequency on itself could also be used as an indicator of emotions with a negative valence (Marcet-Rius et al., 2019). In addition, scores can be allocated to those frames where the whole head of the cow and all the required facial features are visible.

Automated, continuous scoring of cattle disease is interesting in a laboratory setting for e.g. vaccine efficacy trials where consistent 24 h monitoring could increase the accuracy of the detection of the onset of disease. More accurate measurements could lead to a reduction in experimental animals as fewer animals are required to obtain accurate results. One on one extrapolation of the above results to a commercial farm setting is not possible. All animals in the trials were infected at the same time with a known pathogen. In a farm setting, the time of introduction will vary and the identity of the pathogen is probably unknown. Moreover, different pathogens could circulate on any given time. Finally, the age of the animals in this study was homogeneous, which would likely be more heterogeneous in a farm setting. Instead, the results presented here provide a vantage point to a future where the introduction of diseases can be monitored automatically on farms.

Ethics statement

The vaccine efficacy trials were performed in the animal facility of Wageningen Bioveterinary Research, Lelystad, the Netherlands, according to protocols for experimentation with live cattle approved by the Institutional Animal Ethics Committee.

Acknowledgements

We would like to thank Phaedra Eble and Wijbrand Ouweltjes for providing feedback on the document. This study was funded by the Dutch Ministry of Agriculture, Nature and Food Quality (WOT-01-003-

092 en KB-37-003-023).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.prevetmed.2023.105880.

References

Alexandersen, S., Zhang, Z., Donaldson, A.I., Garland, A.J.M., 2003. The pathogenesis and diagnosis of foot-and-mouth disease. J. Comp. Pathol. 129, 1–36. https://doi.org/10.1016/S0021-9975(03)00041-0.

Angelos, J.A., 2020. Infectious Keratoconjunctivitis in Animals [WWW Document]. URL ⟨https://www.msdvetmanual.com/eye-diseases-and-disorders/infectious-keratoconjunctivitis/infectious-keratoconjunctivitis-in-animals⟩ (Accessed 10.6.21).

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Battini, M., Agostini, A., Mattiello, S., 2019. Understanding cows' emotions on farm: are eye white and ear posture reliable indicators? Animals 9, 1–12. https://doi.org/10.3390/ani9080477.

Berckmans, D., 2017. General introduction to precision livestock farming. Anim. Front. 7, 6–11. https://doi.org/10.2527/af.2017.0102.

Boissy, A., Aubert, A., Désiré, L., Greiveldinger, L., Delval, E., Veissier, I., 2011. Cognitive sciences to relate ear postures to emotions in sheep. Anim. Welf. 20, 47–56.

Brown, M.H., Brightman, A.H., Fenwick, B.W., Rider, M.A., 1998. Infectious bovine keratoconjunctivitis: a review. J. Vet. Intern. Med. 12, 259–266. https://doi.org/10.1111/j.1939-1676.1998.tb02120.x.

Descovich, K.A., Wathan, J., Leach, M.C., Buchanan-Smith, H.M., Flecknell, P., Farningham, D., Vick, S.J., 2017. Facial expression: an under-utilized tool for the assessment of welfare in mammals. ALTEX 34, 409–429. https://doi.org/10.14573/altex.1607161.

Donaldson, A.I., Herniman, K.A.J., Parker, J., Sellers, R.F., 1970. Further investigations on the airborne excretion of foor-and-mouth disease virus. J. Hyg. 557–564. https://doi.org/10.1016/0168-1702(90)90084-O.

Gamer, M., Lemon, J., Singh, I.F.P., 2019. irr: Various Coefficients of Interrater Reliability and Agreement.

Gleerup, K.B., 2017. Identifying Pain Behaviors in Dairy Cattle Functions and Effects of Pain 29, pp. 231–239.

Gleerup, K.B., Andersen, P.H., Munksgaard, L., Forkman, B., 2015. Pain evaluation in dairy cattle. Appl. Anim. Behav. Sci. 171, 25–32. https://doi.org/10.1016/j.applanim.2015.08.023.

Goltz, J., 1978. Bluetongue in cattle: a review. Can. Vet. J. 19, 95–98.

Goma, A.A., Pearce, G.P., Uddin, J., Rimon, E., Davies, H., Phillips, C.J.C., 2018. A forced lateralisation test for dairy cows and its relation to their behaviour. Appl. Anim. Behav. Sci. 207, 8–19. https://doi.org/10.1016/j.applanim.2018.06.008.

Hamel, D., Joachim, A., Löwenstein, M., Pfister, K., Silaghi, C., Visser, M., Winter, R., Yoon, S., Cramer, L., Rehbein, S., 2015. Treatment and control of bovine sarcoptic and psoroptic mange infestation with ivermectin long-acting injectable (IVOMEC® GOLD). Parasitol. Res. 114, 535–542. https://doi.org/10.1007/s00436-014-4215-z.

Kadzere, C.T., Murphy, M.R., Silanikove, N., Maltz, E., 2002. Heat stress in lactating dairy cows: a review. Livest. Prod. Sci. 77, 59–91. https://doi.org/10.1016/S0301-6226(01)00330-X.

Lambert, H., Carder, G., 2019. Positive and negative emotions in dairy cows: can ear postures be used as a measure? Behav. Process. 158, 172–180. https://doi.org/10.1016/j.beproc.2018.12.007.

Lambert (Proctor), H.S., Carder, G., 2017. Looking into the eyes of a cow: can eye whites be used as a measure of emotional state? Appl. Anim. Behav. Sci. 186, 1–6. https://doi.org/10.1016/j.applanim.2016.11.005.

Lecy, J.D., Beatty, K.E., 2012. Representative literature reviews using constrained snowball sampling and citation network analysis. SSRN Electron. J. https://doi.org/10.2139/ssrn.1992601.

Mandel, R., Nicol, C.J., Whay, H.R., Klement, E., 2017. Short communication: detection and monitoring of metritis in dairy cows using an automated grooming device. J. Dairy Sci. 100, 5724–5728. https://doi.org/10.3168/jds.2016-12201.

Marcet-Rius, M., Fàbrega, E., Cozzi, A., Bienboire-Frosini, C., Descout, E., Velarde, A., Pageat, P., 2019. Are tail and ear movements indicators of emotions in tail-docked pigs in response to environmental enrichment? Animals 9. https://doi.org/10.3390/ani9070449.

McGuirk, S.M., Peek, S.F., 2014. Timely diagnosis of dairy calf respiratory disease using a standardized scoring system. Anim. Heal. Res. Rev. 15, 145–147. https://doi.org/10.1017/S1466252314000267.

Mendl, M., Burman, O.H.P., Paul, E.S., 2010. An integrative and functional framework for the study of animal emotion and mood. Proc. R. Soc. B Biol. Sci. 277, 2895–2904. https://doi.org/10.1098/rspb.2010.0303.

Moncada, A.C., Neave, H.W., von Keyserlingk, M.A.G., Weary, D.M., 2020. Use of a mechanical brush by dairy cows with chorioptic mange. Appl. Anim. Behav. Sci. 223, 104925 https://doi.org/10.1016/j.applanim.2019.104925.

Müller, B.R., Soriano, V.S., Bellio, J.C.B., Molento, C.F.M., 2019. Facial expression of pain in Nellore and crossbred beef cattle. J. Vet. Behav. 34, 60–65. https://doi.org/10.1016/j.jveb.2019.07.007.

de Oliveira, A.M., Fonseca, A.A., Camargos, M.F., Orzil, L.M., Laguardia-Nascimento, M., Oliveira, A.G.G., Rodrigues, J.G., Sales, M.L., de Oliveira, T.F.P., de Melo, C.B., 2018. Development and validation of rt-qpcr for vesicular stomatitis virus detection (Alagoas vesiculovirus). J. Virol. Methods 257, 7–11. https://doi.org/10.1016/j.jviromet.2018.03.011.

de Oliveira, D., Keeling, L.J., 2018. Routine activities and emotion in the life of dairy cows: Integrating body language into an affective state framework. PLoS One 13, 1–16. https://doi.org/10.1371/journal.pone.0195674.

Peek, S.F., Divers, T.J., 2018. Rebhun's diseases of dairy cattle: Third edition, Rebhun's Diseases of Dairy Cattle: Third Edition.

Proctor, H.S., Carder, G., 2014. Can ear postures reliably measure the positive emotional state of cows? Appl. Anim. Behav. Sci. 161, 20–27. https://doi.org/10.1016/j.applanim.2014.09.015.

Proctor, H.S., Carder, G., 2015. Measuring positive emotions in cows: do visible eye whites tell us anything? Physiol. Behav. 147, 1–6. https://doi.org/10.1016/j.physbeh.2015.04.011.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reefmann, N., Bütikofer Kaszàs, F., Wechsler, B., Gygax, L., 2009. Ear and tail postures as indicators of emotional valence in sheep. Appl. Anim. Behav. Sci. 118, 199–207. https://doi.org/10.1016/j.applanim.2009.02.013.

Rice, J.A., Carrasco-Medina, L., Hodgins, D.C., Shewen, P.E., 2008. Mannheimia haemolytica and bovine respiratory disease. Anim. Heal. Res. Rev. 8, 117–128. https://doi.org/10.1017/S1466252307001375.

Sandem, A.I., Braastad, B.O., 2005. Effects of cow-calf separation on visible eye white and behaviour in dairy cows - a brief report. Appl. Anim. Behav. Sci. 95, 233–239. https://doi.org/10.1016/j.applanim.2005.04.011.

Sandem, A.I., Braastad, B.O., Bøe, K.E., 2002. Eye white may indicate emotional state on a frustration-contentedness axis in dairy cows. Appl. Anim. Behav. Sci. 79, 1–10. https://doi.org/10.1016/S0168-1591(02)00029-1.

Sandem, A.I., Janczak, A.M., Braastad, B.O., 2004. A short note on effects of exposure to a novel stimulus (umbrella) on behaviour and percentage of eye-white in cows. Appl. Anim. Behav. Sci. 89, 309–314. https://doi.org/10.1016/j.applanim.2004.06.011.

Schmied, C., Waiblinger, S., Scharl, T., Leisch, F., Boivin, X., 2008. Stroking of different body regions by a human: effects on behaviour and heart rate of dairy cows. Appl. Anim. Behav. Sci. 109, 25–38. https://doi.org/10.1016/j.applanim.2007.01.013.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 7881.

Smith, B.I., Risco, C.A., 2005. Management of periparturient disorders in dairy cattle. Vet. Clin. N. Am. - Food Anim. Pr. 21, 503–521. https://doi.org/10.1016/j.cvfa.2005.02.007.

Stangaferro, M.L., Wijma, R., Caixeta, L.S., Al-Abri, M.A., Giordano, J.O., 2016a. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part I. Metabolic and digestive disorders. J. Dairy Sci. 99, 7395–7410. https://doi.org/10.3168/jds.2016-10907.

Stangaferro, M.L., Wijma, R., Caixeta, L.S., Al-Abri, M.A., Giordano, J.O., 2016b. Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part III. Metritis. J. Dairy Sci. 99, 7422–7433. https://doi.org/10.3168/jds.2016-11352.

Starič, J., Križanec, F., Zadnik, T., 2008. Listeria monocytogenes keratoconjunctivitis and uveitis in dairy cattle. Bull. Vet. Inst. Pulawy 52, 351–355.

Tamioso, P.R., Rucinque, D.S., Taconeli, C.A., da Silva, G.P., Molento, C.F.M., 2017. Behavior and body surface temperature as welfare indicators in selected sheep regularly brushed by a familiar observer. J. Vet. Behav. Clin. Appl. Res 19, 27–34. https://doi.org/10.1016/j.jveb.2017.01.004.

van Dierendonck, M.C., van Loon, J.P.A.M., 2016. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): a validation study. Vet. J. 216, 175–177. https://doi.org/10.1016/j.tvjl.2016.08.004.

van Loon, J.P.A.M., Van Dierendonck, M.C., 2015. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): a scale-construction study. Vet. J. 206, 356–364. https://doi.org/10.1016/j.tvjl.2015.08.023.

Welfare Quality, 2009. Welfare Quality Assessment protocol for cattle, Welfare Quality® Assessment Protocol for Cattle.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686.

Yates, W.D.G., 1982. A review of infectious bovine rhinotracheitis, shipping fever pneumonia and viral-bacterial synergism in respiratory disease of cattle. Can. J. Comp. Med. 46, 225–263.