A six-step process to explore facial expressions performances to detect pain in dairy cows with lipopolysaccharide-induced clinical mastitis

L. Ginger, L. Aubé, Dorothee Ledoux, M. Borot, C. David, Matthieu Bouchon, M. Leach, Denys Durand, Alice de Boyer Des Roches

Applied Animal Behaviour Science 264 (2023) 105951

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

A six-step process to explore facial expressions performances to detect pain in dairy cows with lipopolysaccharide-induced clinical mastitis

L. Ginger [a,1,2], L. Aubé [a,b,1,3], D. Ledoux [a,b,4], M. Borot [c,5], C. David [c], M. Bouchon [d], M. Leach [e,6], D. Durand [a,7], A. de Boyer des Roches [a,b,*,8]

[a] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR Herbivores, 63122 Saint-Genès-Champanelle, France
[b] Chaire Bien-Etre Animal, VetAgro Sup, 1 Avenue Bourgelat, 69280 Marcy-l'Etoile, France
[c] Université de Lyon, VetAgro Sup, 1 Avenue Bourgelat, 69280 Marcy-l'Etoile, France
[d] INRAE, Herbipôle, 63122 Saint-Genès-Champanelle, France
[e] School of Natural and Environmental Science, Newcastle University, Newcastle-upon-Tyne, United Kingdom

ARTICLE INFO

Keywords:
Cattle
Udder inflammation
Facial expression
Pain assessment

ABSTRACT

This study aimed to explore the performances of Facial Action Units (**FAUs**) to detect pain in cows under lipopolysaccharide (**LPS**) challenge though a six-step process based on expression, selectivity (construct validity, responsiveness), intra- and inter-observer reliability, and measurement error (specificity, **Sp**; sensitivity, **Se**). Twenty-seven cows received an intra-mammary infusion of 25 μg *E. coli* LPS in one healthy udder quarter. Then, 14 cows received a placebo (LPS cows) and 13 cows received 3 mg/kg BW ketoprofen i.m. (LPS+NSAID cows). Each cow's face was video-recorded for 40 s at three time points before (T-21 = 11:30 h, T-19 = 13:30 h, T-16 = 16:30 h) and after (T3 = 11:30 h, T5 = 13:30 h, T8 = 16:30 h) infusion. Three trained observers scored the duration, frequency or presence/absence of 43 FAUs on 40 second video segments. We kept only the selective FAUs (construct validity) and transformed them into binary variables (below / above a threshold determined by ROC curves and Youden index). Intra- and inter-observer reliability were assessed by percentage of agreement (**PA**) and Fleiss' kappa (**k**). We calculated Sp and Se. The process therefore consisted in a waterfall method with expression, selectivity, intra-observer reliability being an eliminative step, while inter-observer reliability, Sp and Se were not. FAUs were kept if expressed (>5% of the videos), and kept for intra-observer reliability if PA≥ 75% and k ≥ 0.41. Two too rarely expressed FAUs were excluded. LPS infusion induced changes in 7 FAUs of orbital, auricular, and mouth-and-muzzle regions. Compared to before challenge (T-19), at T5, LPS cows spent significantly less time with 'muzzle in motion' (P = 0.045), tended to display more 'nostril dilation' (P = 0.097), spend more time with 'motionless muzzle' (P = 0.068) and less time in 'ear: position 8′ (i.e. backwards /central/pinna to the side) (P = 0.057). At T8, LPS cows spent significantly less time with 'eye open' (P = 0.036) and tended to less frequently display 'eye blinking' (P = 0.071) and 'eye movements' (P = 0.091) compared to T-16. Four of these 7 FAUs ('eye open', 'motionless muzzle', 'muzzle in motion', 'nostril dilatation') satisfied all following steps of the process except sensitivity. Two other FAUs ('eye blinking', 'eye movements') satisfied intra- and inter-observer reliability; depending on the time point considered they were either sensitive or specific but not both simultaneously. The last FAU ('ear: position 8′) satisfied intra-observer reliability and sensitivity but not inter-observer reliability nor specificity. This study identified 7 FAUs as potential candidate for detecting mild pain associated with induced inflammatory mastitis in dairy cows.

* Correspondence to: UMR 1213 Herbivores INRA–VetAgro Sup–Campus Vétérinaire de Lyon, 1 Avenue Bourgelat, 69280 Marcy-l'Etoile, France.
E-mail address: alice.deboyerdesroches@vetagro-sup.fr (A. de Boyer des Roches).
[1] Joint first authors (co-first authorship)
[2] ORCID https://orcid.org/0000-0002-5199-5517
[3] ORCID https://orcid.org/0000-0002-6174-3743
[4] ORCID https://orcid.org/0000-0001-5448-5689
[5] ORCID https://orcid.org/0000-0002-7175-7233
[6] ORCID https://orcid.org/0000-0002-7148-0158
[7] ORCID https://orcid.org/0000-0001-8381-8999
[8] ORCID https://orcid.org/0000-0002-6903-7456

https://doi.org/10.1016/j.applanim.2023.105951
Received 12 December 2022; Received in revised form 4 May 2023; Accepted 11 May 2023
Available online 12 May 2023

1. Introduction

Facial expressions are defined as changes in the face or the contraction/relaxation of face-muscle movements in response to emotional stimuli (Williams, 2002; Mogil et al., 2020). Facial expression assessment provides several advantages for use in behavioral research (Sotocinal et al., 2011; Dalla Costa et al., 2014; Descovich et al., 2017). First, humans have a natural predisposition to focus on face (Williams, 2002; Leach et al., 2011). Second, facial expressions can be assessed by direct observations and require no special equipment or physical contact, thus avoiding additional stress for the animal and ensuring security for the observer.

Over the last decade, studies have investigated the emotional component of pain expressed through scales exclusively based on modifications of facial expressions: 'grimace scales'. These scales were first developed in humans (Mogil et al., 2020), then in laboratory animals, starting with mice (Langford et al., 2010) followed by rats (Sotocinal et al., 2011) and rabbits (Banchi et al., 2020), and in other species: cats (Evangelista et al., 2019), ferrets (Reijgwart et al., 2017), horses (Dalla Costa et al., 2014; Gleerup et al., 2015b), foals (Lanci et al., 2022), donkeys (Orth et al., 2020), sheep (Häger et al., 2017) and lambs (Guesgen et al., 2016), pigs (Di Giminiani et al., 2016) and cattle (Gleerup et al., 2015a; Müller et al., 2019; Yamada et al., 2021). These studies have used various pain contexts including e.g. castration (Dalla Costa et al., 2014; Yamada et al., 2021), mastitis or footrot (McLennan et al., 2016), tail docking (Di Giminiani et al., 2016; Guesgen et al., 2016) and hot branding (Müller et al., 2019). These grimace scales generally comprise four regions of the face, i.e. orbital, nose/muzzle, mouth, and auricular. Then, for each region, a set of units of actions are defined that each describe a specific facial expression. These units of actions are called Facial Action Units (**FAUs**).

To be considered valid for assessing pain in animals, a measure has to satisfy several performance-characteristics (Mogil et al., 2020; Knierim et al., 2021), such as those from the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines (Mokkink et al., 2016; Tomacheuski et al., 2023). Some of the performances characteristics are: i) selectivity (i.e. ability to quantify what it is supposed to quantify) checked by construct validity (i.e. discrimination between painful and pain-free animals) and responsiveness (i.e. discrimination between animals with vs without analgesic intervention), ii) Reliability, assessed through intra-observer reliability ('repeatability'), and inter-observer reliability ('reproducibility'), and iii) measurement error assessed through specificity (i.e. ability to correctly detect pain-free animals, **Sp**) and sensitivity (i.e. ability to correctly detect animals in pain, **Se**).

Most studies on facial expression in a pain context have explored selectivity but left reliability and measurement error quite incomplete, especially in cattle (reviewed in Evangelista et al., 2022 and Tomacheuski et al., 2023). Most authors have reported a high inter-observer reliability for the entire grimace scale, e.g. intraclass correlation coefficient (ICC) = 0.90 in mice (Langford et al., 2010), ICC = 0.92 in horses (Dalla Costa et al., 2014), and ICC = 0.86 in sheep (McLennan et al., 2016). The few studies exploring each FAU reported low-to-high inter-observer reliability (high in rats with ICC ranging from 0.86 to 0.96 (Sotocinal et al., 2011); low to high in foals with Fleiss Kappa ranging from 0.24 to 1.0 (Lanci et al., 2022); moderate to high in horses and sheep with ICC ranging from 0.58 to 0.97 (Dalla Costa et al., 2014) and from 0.63 to 0.90 (McLennan et al., 2016), coefficients being interpreted as suggested by Landis and Koch (1977) and Koo and Li (2016). Both intra- and inter-observer reliability were explored together in foals (Lanci et al., 2022) or cats (Evangelista et al., 2019; Luna et al., 2022), but not in cattle(Gleerup et al., 2015a; Müller et al., 2019; Yamada et al., 2021). Finally, to our knowledge, measurement error (Sp and Se) have only been explored for some entire grimace scales (e.g. in McLennan et al., 2016, Luna et al., 2022) but not for each FAU. There is therefore currently a lack of knowledge regarding all performance-characteristics of FAUs for detecting pain in animals.

To date, four studies have used facial expressions to assess pain in cattle (Gleerup et al., 2015a; de Boyer des Roches et al., 2017; Müller et al., 2019; Yamada et al., 2021). Two of them (Gleerup et al., 2015a; de Boyer des Roches et al., 2017) aimed to create a methodology and not carry out an appropriate validation. The two remaining studies (Müller et al., 2019; Yamada et al., 2021) only used a partially validated tool to investigate a painful situation. It is therefore necessary to comprehensively validate a set of FAUs to reliably detect pain in cattle.

Mastitis is a widely used model for pain research, as it can be experimentally induced, controlled and modulated. Experimental infusion of lipopolysaccharide (**LPS**), a pro-inflammatory immunogenic cell-wall component of Gram-negative bacteria, into the mammary gland provokes a local inflammation and a systemic response similar to early clinical phases of mastitis. The responses can be also modulated by administration of non-steroidal anti-inflammatory drugs (**NSAIDs**), which have analgesic and antipyretic effects (Zimov et al., 2011; Fitzpatrick et al., 2013).

The aim of this study was to find potential Facial Action Units (**FAUs**) that could be included in future pain grimace scale in cattle. To this aim, we explored FAUs to detect pain in dairy cows with udder inflammation through a six-steps process adapted from COSMIN methodology (Evangelista et al., 2022; Tomacheuski et al., 2023): (1) FAU's expression (is the FAU expressed by cows?), (2) selectivity (through construct validity and responsiveness), (3) intra and (4) inter-observer reliability, measurement error i.e. (5) specificity and (6) sensitivity, using a kind of waterfall method (at some steps the FAUs could be eliminated from the process). We used an *E.coli* LPS mammary challenge model with or without NSAID pain relief (Ginger et al., 2023).

2. Materials and methods

This study was conducted as part of a larger study on behavioral pain assessment in dairy cows. Detailed information on animals, protocols and procedures is comprehensively described elsewhere (Ginger et al., 2023).

2.1. Ethics statement

The study was conducted from February to March 2019 at the 'Herbipôle' multidisciplinary experimental research platform (Herbipôle, INRAE, 2018), an upland ruminant farming systems research facility (doi:10.15454/1.5572318050509348E12) located at Marcenat in France. All experimental protocols and procedures were carried out with the approval of the local CEMEA Auvergne institutional animal care and use committee (CE-05092.01, APAFIS agreement #2015043014541577) and conducted in full compliance with all applicable provisions established by European Directive 2010/63/EU. All procedures were applied by trained staff members who performed the experiment in accordance with all relevant named guidelines and regulations. The study was carried out in accordance with ARRIVE guidelines. All animals used in this study were handled in strict adherence to good clinical practices, and every effort was made to minimize suffering. Endpoints were defined before the start of the experiment: any cows showing any signs of sickness or distress during the experiment were examined by a veterinarian, and removed from the study if they crossed the threshold limit of a rectal temperature above 42.5 °C for four consecutive hours together with a score above 12 on the de Boyer des Roches et al. (2017) grid. After the trial, the cows returned to the herd of the experimental unit. Throughout the trial period and up to 28 days later, their milk will was diverted from collection for human consumption.

2.2. Animals, housing, and feeding

The study used 28 Holstein dairy cows (mean age 3.2 ± SD 0.2

years). The cows were primiparous, at 128.9 ± 15.8 (between 100 and 163) days in milk, with a body condition score of 2.5–3 (scored from 1 to 5; Edmonson et al., 1989)) and either pregnant or in the luteal phase. All cows were kept together in a loose-housing cubicle barn (244 m^2) with 28 cubicles, 28 self-locking barriers, a watering trough, one salt lick, and an automatic rotating brush (DeLaval SCB, DeLaval, Sweden). The building was lit from 05:00 h to 20:00 h, and dimmed lights were left on at night. Cows were fed with a total mixed ration, adapted to dietary requirements for the lactation period, containing hay (11.3 kg DM per cow), beet molasses (0.4 kg DM per cow), haylage (5.8 kg DM per cow), concentrate (2.7 kg DM per cow, with nitrogen corrector and energy corrector) and minerals. The mixed ration was distributed once per day (at 10:00 h) and pushed back toward the cows three times a day (at 13:00 h, 16:00 h and 22:00 h). The cows were milked twice a day (around 07:30 h and 16:30 h) by two experienced stockpersons in a (2 ×14) milking parlor adjacent to the barn.

2.3. Experimental design

The experiment examined the effects of an experimental *E. coli* LPS infusion in the mammary gland and the effects of intramuscular injection of a **NSAID**, ketoprofen (Ketofen® 10%, 3 mg/kg BW; CEVA Santé Animale, Libourne, France) or placebo (0.9% NaCl, 3 mL/100 kg, BIOLUZ, St-Jean-de-Luz, France) on dairy cow FAUs. The 28 cows were randomly assigned to the following treatments (see Ginger et al. 2023 for details): intra-mammary LPS challenge followed by placebo, i.e. intramuscular saline injection (**LPS**, n = 14 cows), and intra-mammary LPS challenge followed by intramuscular ketoprofen injection (**LPS+NSAID**, n = 14 cows; see below).

2.4. Pain model

2.4.1. Intra-mammary Escherichia coli LPS challenge and NSAID treatment

Before challenge, somatic cell count (SCC) was measured (Log10SCC at T-1 was 4.04 [3.90– 4.18], (Ginger et al., 2023) and bacteriological analyses on milk were performed to check the absence of intra-mammary infection in all udder quarters of the 28 cows. For each cow, one quarter was selected to be challenged based on the following criteria: i) priority was given to hindquarters because they were more easy to access in the herringbone milking parlor, and allowed comparisons with Fitzpatrick et al. (2013)' study, ii) with the lowest cell count (i.e. SCC < 50,000 cells/mL in the quarter at the evening milking on the day prior to challenge, i.e. T-16), and iii) with a negative bacteriological result (see Ginger et al. 2023 for details). The experimental intra-mammary LPS infusion was performed after the morning milking on the day of challenge at T0 (Fig. 1). For each cow, the quarter selected was infused with 25 μg ultra-pure LPS from *E. coli* O111 (tlrl-3pelps, InVivogen, Toulouse, France) diluted in 2 mL of a sterile solution of DPBS containing 0.5% (w/v) sterile bovine serum albumin solution for cell culture (BSA, Sigma). The cows were then returned to their home pen and headlocked. There, the cows received an intramuscular injection of either saline solution (0.9% NaCl, 3 mL/100 kg, BIOLUZ, St-Jean-de-Luz, France; LPS group) or ketoprofen (Ketofen® 10%, 3 mg/kg; CEVA Santé Animale, Libourne, France; LPS+NSAID group) injected swiftly after the challenge procedure. Detailed experimental procedures for LPS infusion and treatments injections can be found in Ginger et al. (2023).

2.4.2. Validation of the pain model

Infusion of *E.coli* LPS in the udder induced an acute udder inflammation demonstrated by a significant increase in rectal temperature, heart rate, respiratory rate and interleukins (IL-6, IL-1b, IL-8) at 8 h post infusion (**hpi**) and in somatic cells in milk from 8 hpi up to 48 hpi (Ginger et al., 2023). LPS cows experienced pain of short duration (from 3 hpi up to 8 hpi) demonstrated by increased plasma cortisol at 3 hpi and 8 hpi, increased cortisol in milk at 8 hpi, and behavioral changes in the barn and at milking: more cows pressed their tail against their udder at 5 hpi, lifted their foot at milking at 8 hpi, and decreased their feeding and ruminating activities at 3 hpi and 5 hpi (Ginger et al., 2023).

Intramuscular injection of ketoprofen (LPS+NSAID cows) significantly decreased (i.e. improved) endocrine and clinical parameters: compared to LPS cows, LPS+NSAID cows had lower plasma cortisol levels at 3 hpi and lower rectal body temperature at 8 hpi, and higher heart rate at 1 hpi and 32 hpi. They also recovered normal feeding/ruminating activity, ear posture, and body posture at 5 hpi (Ginger et al., 2023).

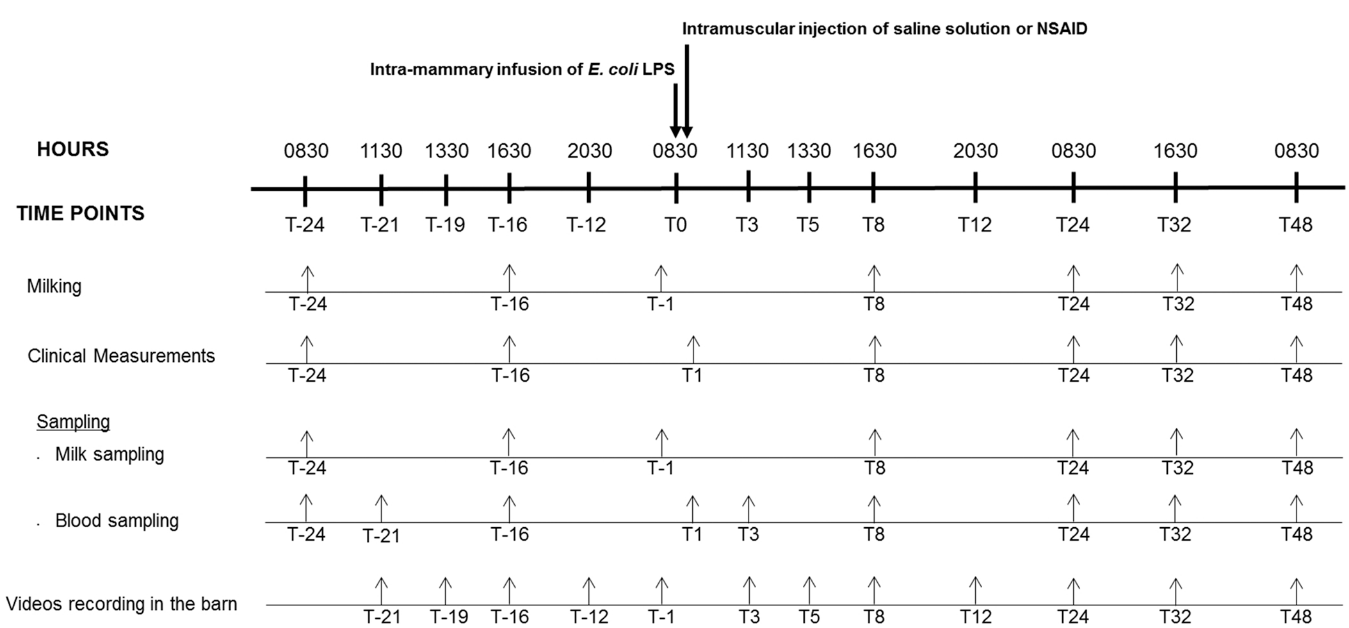


Fig. 1. Experimental protocol used to examine the effects of experimental 25 μg *E. coli* LPS infusion in the mammary gland and the effects of NSAID (ketoprofen, Ketofen® 10%, 3 mg/kg; CEVA Santé Animale Libourne, France) (LPS+NSAID cows, n = 13) or placebo (saline solution, 0.9% NaCl, 3 mL/100 kg, BIOLUZ, St-Jean-de-Luz, France) (LPS cows, n = 14) injection in 27 Holstein-Friesian dairy cows: milkings, clinical measurements, milk and blood sampling, and in-barn video-recording.

2.5. Video recording

No experimental cow was in heat at the moment of observation. One cow showed signs of severe lameness (score 5 on Sprecher's scale; Sprecher et al., 1997) before the scheduled challenge and was therefore excluded from the experiment before the challenge. No cow reached an endpoint. The final number of cows analyzed was therefore 27, i.e. 14 LPS cows and 13 LPS+NSAID cows (Ginger et al., 2023).

The 27 cows were individually video-recorded at twelve time points: T-21 (11:30 h), T-19 (13:30 h), T-16 (16:30 h), T-12 (20:30 h), T-1 (6:00 h), T3 (11:30 h), T5 (13:30 h), T8 (16:30 h), T12 (20:30 h), T24 (8:30 h), T32 (16:30 h), T48 (8:30 h) (Fig. 1) (324 videos in total, 168 of LPS cows and 156 of LPS+NSAID cows, Table S1). Each video recording lasted three minutes and was recorded with a camcorder (Panasonic HC-VX870 4 K, Japan) by a trained researcher assisted by a MSc student, both of whom were blinded to cow treatment. The researcher stood still at the boundary of the pen, at 5–8 m to film the focal cow's whole body for two minutes, then zoomed in on the cow's head for one minute. Video-recordings were always carried out before any interventions on animals (e.g. blood sampling, milking).

2.6. Video used for analysis

A PhD student, blinded to cow treatment and time point, selected 40-second sequences from the 1-minute zoom on the cow's face from each film. Any poor-quality video (e.g. head not visible, low luminosity, bad angle of view) was removed from the video set (Table S1). The initial video set included 315 40-second video sequences: 165 videos of LPS cows and 150 videos of LPS+NSAID cows (Table S1).

The number of videos used varied according to the step of the six-step process. For FAU's expression (step 1) and selectivity (step 2), 156 videos were used: 83 videos of LPS cows and 73 videos of LPS+NSAID cows (video set 1). They were chosen according to the pain status of the animals, with 27 videos for each time point at which the cows were pain-free (T-21, T-19, T-16); and 27 videos for each time point at which the cows were considered to be experiencing pain (T3, T5 and T8) (Ginger et al., 2023) (Table S1). Nevertheless, one video from LPS cows before challenge and five videos from LPS+NSAID cows before challenge were removed from the video set (Table S1), due to several reasons (e.g. head of the cow not visible because the cow was eating, or against a penmate, or due to a recording problem). For intra-observer reliability (Step 3), 21 videos were used (video set 2). They were randomly chosen from the entire video set. Among them, 8 came from LPS cows and 9 came from LPS+NSAID cows; 6 different videos had been recorded before challenge (3 were recorded at T-19, 3 were recorded at T-1), 15 had been recorded after challenge: 6 different videos had been recorded during the acute phase of pain (3 were recorded at T3, 3 were recorded at T8), and 9 different videos had been recorded after (3 were recorded at T24, 3 were recorded at T32, and 3 were recorded at T48). For inter-observer reliability (step 4), specificity (step 5) and sensitivity (Step 6), 156 videos were used: 83 videos of LPS cows and 73 videos of LPS+NSAID cows, with 24 videos at T-21, 26 videos at T-19 and 25 videos at T-16, 27 videos at T3, 27 videos at T5 and 27 videos at T8 (video set 1).

2.7. Identification of potential FAUs for detecting pain in cows

Based on the literature and expert consultation, FAUs were selected by a PhD student who viewed the 156 videos (video set 1) to check the feasibility of observation.

An ethogram was developed to describe each FAU in the three regions (orbital region, muzzle + mouth region, and auricular region) defining 29 initial FAUs (Table 1, Fig. 2). FAUs were scored either in seconds (i.e. time spent by the cow in the item, 'state' FAUs), in presence/absence, or in number (i.e. frequency within the 40 s of observation, 'event' FAUs). In addition to the initial FAUs presented in Table 1, some FAUs were combined to allow comparison with previous work (Mialon et al., 2012; Gleerup et al., 2015a; b; McLennan et al., 2016; Lambert and Carder, 2019; Müller et al., 2019; Yamada et al., 2021). For example, the FAU "ear backward" in Table 1 was a composite of five FAUs: ear position 5 + ear position 6 + ear position 7 + ear position 8 + ear position 9. The final cow pain facial expression scale comprised 43 FAUs (Table 1).

2.8. Observers and training

Three observers (**Obs1**, **Obs2** and **Obs3**), blinded to cow treatment, scored the 156 videos of video set 1. Obs1 (man) was a technician in animal behavioral science with 22 years of experience in video scoring, and Obs2 and Obs3 (women) were two final-year veterinary students specialized in cattle medicine.

Before starting video coding, the observers followed a training program covering acquisition of each FAU (Obs1, 2 and 3) (Table 1), familiarization with the scoring software (The Observer XT 14, Noldius, Wageningen, The Netherlands) (Obs2 and Obs3), and a practice scoring session (Obs1, 2 and 3).

2.9. Development of FAUs as pain indicators and statistical analyses

In order to determine which FAUs were valid for assessing pain associated with udder inflammation in cows, a six-steps process was defined (Fig. 3). It comprised six steps adapted from COSMIN methodology (Evangelista et al., 2022; Tomacheuski et al., 2023): (1) FAU expression by cows, (2) selectivity, (3) intra-observer reliability (or 'repeatability'), (4) inter-observer reliability (or 'reproducibility'), measurement error i.e. (5) specificity, and (6) sensitivity. Each FAU was studied step by step, and if it did not satisfy the performance requirements for selectivity (through construct validity) or intra-observer reliability, it was not further explored (Fig. 3). Here we report an analysis of the facial expressions of the cows at six time points: three pre-challenge time points (T-21 = 11:30 h, T-19 = 13:30 h, T-16 = 16:30 h) at which the cows were pain-free, and three post-challenge time points (T3 = 11:30 h, T5 = 13:30 h, T8 = 16:30 h) at which the cows were considered to be experiencing pain (Ginger et al., 2023).

2.9.1. Expression of FAUs by cows

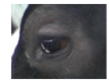
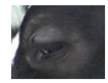
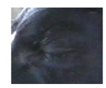
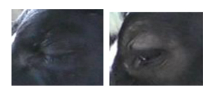
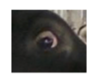
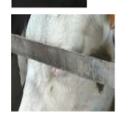
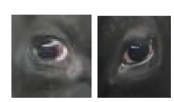
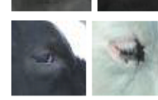
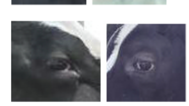
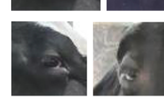
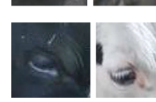
To evaluate whether a FAU was relevant to be included in the next step of the waterfall process, we checked whether it was expressed in at least 8 videos among the set of 156 videos (i.e. LPS cows and LPS + NSAID cows coded by Obs 1 at the T-21 (11:30 h), T-19 (13:30 h), T-16 (156:30 h), T3 (11:30 h), T5 (13:30 h) and T8 (16:30 h). This corresponds to 5% of the videos. All FAUs that were expressed in less than 5% of the videos were considered too rare and were thus eliminated from the next steps of the process (Fig. 3).

2.9.2. Selectivity of FAUs to assess pain

2.9.2.1. FAU's evaluation after LPS infusion: construct validity. To evaluate changes in each FAU after LPS infusion in the udder, we used the videos from LPS cows scored by Obs1. Each cow was her own control. To eliminate the circadian rhythm effect on cows' behavior (Veissier et al., 2017), we compared changes in each FAU at the same hour of the day before vs. after challenge (i.e. 11:30 h: T-21 vs. T3, 13:30 h: T-19 vs. T5, 16:30 h: T-16 vs. T8). Data were graphically checked for normality. Because data did not follow a normal distribution, duration or frequency of each FAU before and after challenge were compared using a Wilcoxon rank sum test. Proportion of LPS cows assigned to a 'presence' or 'absence' modality before vs. after challenge was compared using a McNemar test (Table 2).

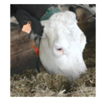
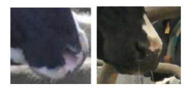
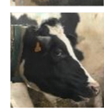
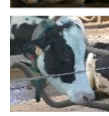
FAUs that differed significantly (P-value < 0.05) or tended to differ ($0.05 \leq$ P-value < 0.1) at least one time point (i.e. before vs. after LPS infusion) were considered as selective for detecting pain and thus

Table 1
Ethogram of facial expressions.

Head region	Facial action unit[2]	Coding unit	Description	Illustration	References
Orbital	Eye open	Seconds	Eyeball visible in its entirety, eyelid visible and does not fall down, and sclera visible in part. The cow can blink.		Langford et al. (2010);Sotocinal et al. (2011); Keating et al. (2012);Dalla Costa et al. (2014);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016);Müller et al. (2019);Yamada et al. (2021);Lanci et al. (2022)
	Half-closed eye	Seconds	Eyeball not visible in its entirety, eyelid is visible, half-closed and not covering the entire eyeball		Langford et al. (2010);Sotocinal et al. (2011); Keating et al. (2012);Dalla Costa et al. (2014);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016);Müller et al. (2019);Yamada et al. (2021);Lanci et al. (2022)
	Closed eye	Seconds	Eyeball not visible and eyelid fully visible and covers the entire eyeball		Langford et al. (2010);Sotocinal et al. (2011); Keating et al. (2012);Dalla Costa et al. (2014);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016);Müller et al. (2019);Yamada et al. (2021);Lanci et al. (2022)
	Eye closure: half-closed eye + closed eye	Seconds	Eyeball is partially or not visible; eyelid is fully visible and covers the entire eyeball or partially visible, half closed and not covering the entire eyeball		Müller et al. (2019);Yamada et al. (2021)
	Wide-open eye	Seconds	Entire eyeball visible, eyelid almost not visible and sclera visible for the most part		Lanci et al. (2022)
	Eyes not visible	Seconds	Not visible, either because the cow turns its head, or because an external element (barrier, strap, hay, etc.) prevents the eye from being seen OR because the camera is far from the cow, the angle of the shot is too narrow, the brightness is too low		
	Eye blinking	Numbers	The eyelid drops and rises immediately. Duration of the movement is less than one second.		Gleerup et al. (2015a)
	Eye contraction	Numbers	The eyelid drops and rises. We can see it contracting. The duration of the movement is more than one second.		
	Eye movements: eye blinking + eye contraction	Numbers	The eyelid drops and rises without notion of duration.		Gleerup et al. (2015a)
	Tears	Numbers	A tear flows directly from the open eye.		
	Humid eye	Presence Absence Not visible	Distinct tear film, more so than in a normal situation No visible tear film		
	Eye discharge	Presence Absence Not visible	Abnormal and excessive discharge from the medial corner of the eye, more or less humid No flow or presence of old flow		
	Hollow eye	Presence Absence Not visible	Sinking of the eye in its orbit. Appearance of a hollow above the upper eyelid and below the lower eyelid No sinking of the eye in its orbit		Gleerup et al. (2015a)
	Swelling of the eyebrow	Presence Absence Not visible	Swelling of the arch (increase in volume) Normal-sized arch and no abnormal swelling		Gleerup et al. (2015a);Di Giminiani et al. (2016)
	Striations of the eyebrow arch	Presence Absence Not visible	Vertical striations in the medial part of the arch of the eyebrow Smooth brow bone		Gleerup et al. (2015b)
Muzzle and mouth	Motionless muzzle	Seconds	Muzzle is motionless, and is immobile from 2 s without movement.		
	Mobile muzzle	Seconds	Mouth in movement: either it rolls up, or the nostrils dilate or any other movement		
	Mobile-eating muzzle	Seconds	Muzzle in motion because the cow is eating		Gleerup et al. (2015b)
	Mobile-ruminating muzzle	Seconds	Mouth is in motion. We can see a movement of the lower jaw going from left to right and vice versa.		Gleerup et al. (2015b)
	Muzzle in motion due to eating or ruminating: Mobile-eating muzzle	Seconds	All movements of the muzzle and mouth due to eating or ruminating activities		

(*continued on next page*)

Table 1 (*continued*)

Head region	Facial action unit[2]	Coding unit	Description	Illustration	References
	+ Mobile-ruminating muzzle				
	Muzzle not visible	Seconds	Not visible, either because the cow turns its head or because an external element (barrier, strap, hay, etc.) prevents the eye from being seen		
	Vertical movements of muzzle	Numbers	Vertical movement of the muzzle from the bottom to the top, with a curling movement.		
	Nostril dilations	Numbers	Increase in volume of the nostrils		Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Müller et al. (2019);Lanci et al. (2022)
	Muzzle discharge	Presence Absence Not visible	Mucous, purulent or mixed discharge Moist but no flow		Gleerup et al. (2015b)
	Prominent muzzle muscle or veins	Presence Absence Not visible	More prominent facial vein and/or muzzle muscles or more prominent vein and/or muscle contours Smooth chanfron		Dalla Costa et al. (2014);McLennan et al. (2016);Müller et al. (2019);Yamada et al. (2021)
Auricular[1]	Ear: position 1 (forward/ center/forward)	Seconds	Forward according to cranio-caudal axis, center according to dorso-ventral axis, and forward according to orientation of the pinna		Sotocinal et al. (2011);Gleerup et al. (2015b);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019)
	Ear: position 2 (middle/ upward/forward)	Seconds	Middle according to cranio-caudal axis, upward according to dorso-ventral axis, and forward according to orientation of the pinna		Sotocinal et al. (2011);Gleerup et al. (2015b);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019)
	Ear: position 3 (middle/ center/forward)	Seconds	Middle according to cranio-caudal axis, center according to dorso-ventral axis, and forward according to orientation of the pinna		Sotocinal et al. (2011);Gleerup et al. (2015b);Di Giminiani et al. (2016); Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019)
	Ear: position 4 (middle/ downward/forward)	Seconds	Middle according to cranio-caudal axis, downward according to dorso-ventral axis, and forward according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011); Keating et al. (2012);Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019)
	Ear: position 5 (backward/ upward/backward)	Seconds	Backward according to cranio-caudal axis, upward according to dorso-ventral axis, and backward according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011);Dalla Costa et al. (2014);Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019);Müller et al. (2019)
	Ear: position 6 (backward/ upward/to the side)	Seconds	Backward according to cranio-caudal axis, upward according to dorso-ventral axis, and to the side according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011);Dalla Costa et al. (2014);Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019);Müller et al. (2019)
	Ear: position 7 (backward/ upward/downward)	Seconds	Backward according to cranio-caudal axis, upward according to dorso-ventral axis, and downward according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011);Dalla Costa et al. (2014);Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019);Müller et al. (2019)
	Ear: position 8 (backward/ center/forward)	Seconds	Backward according to cranio-caudal axis, center according to dorso-ventral axis, and to the side according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011);Dalla Costa et al. (2014);Gleerup et al. (2015a);Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan et al. (2016); Lambert and Carder (2019);Müller et al. (2019)
	Ear: position 9 (backward/ downward/to the side)	Seconds	Backward according to cranio-caudal axis, downward according to dorso-ventral axis, and to the side according to orientation of the pinna		Langford et al. (2010);Sotocinal et al. (2011); Keating et al. (2012);Dalla Costa et al. (2014);Gleerup et al. (2015a); Gleerup et al. (2015b);Di Giminiani et al. (2016);Guesgen et al. (2016);McLennan

(*continued on next page*)

Table 1 (*continued*)

Head region	Facial action unit[2]	Coding unit	Description	Illustration	References
					et al. (2016); Lambert and Carder (2019); Müller et al. (2019)
	Ear: position 2 + position 3	Seconds	Position 2: middle/upward/forward Position 3: middle/center/forward		
	Ear: position 5 + position 6 + position 7	Seconds	Position 5: backward/upward/backward Position 6: backward/upward/to the side Position 7: backward/upward/downward		
	Ear: position 8 + position 9	Seconds	Position 8: backward/center/forward Position 9: backward/downward/to the side		
	Ear in an intermediate position	Seconds	Ear: position 2 + position 3 + position 4		Mialon et al. (2012)
	Ear backward	Seconds	Ear: position 5 + position 6 + position 7 + position 8 + position 9		Mialon et al. (2012);Gleerup et al. (2015b); Lambert and Carder (2019); Müller et al. (2019)
	Ear upward	Seconds	Ear: position 2 + position 5 + position 6 + position 7		
	Ear in median position	Seconds	Ear: position 1 + position 3 + position 8		
	Ear downward	Seconds	Ear: position 4 + position 9		Gleerup et al. (2015b); Lambert and Carder (2019)
	Ear with pinna toward the front	Seconds	Ear: position 1 + position 2 + position 3 + position 4 + position 8		McLennan et al. (2016)
	Ear with pinna toward the side	Seconds	Ear: position 6 + position 9		McLennan et al. (2016)
	Ear backward and/or downward	Seconds	Ear: position 4 + position 5 + position 6 + position 7 + position 8 + position 9		McLennan et al. (2016)
	Ear not visible	Seconds	The ear is not visible because an external element prevents the ear from being seen (barrier, strap, etc.)		

[1] Ear positions were described according to three axes: horizontal, vertical, and pinna axis. The horizontal axis is determined by the axis between the head and the tail of the cow: ear can be oriented forward (ear tip pointed toward muzzle), backward (ear tip pointed toward tail), or in the middle (ear tip is perpendicular to the horizontal axis). The vertical axis is determined by the axis between ground and sky: ear can be oriented upward (ear tip pointed toward sky), downward (ear tip pointed toward ground), or in center (ear tip is central/perpendicular to the vertical axis). The pinna axis is determined by the orientation of the pinna when the observer is standing in front of the cow: pinna can be oriented forward (entire pinna is visible), backward (pinna is not visible), to the side (pinna is partially visible and oriented parallel to the horizontal axis) or downward (pinna is partially visible and oriented toward the ground).

[2] The FAUs 'eye open', 'half-closed eyes', 'eye closed', 'eyes wide open' and 'eye not visible' are mutually exclusive; the FAUs 'motionless muzzle', 'mobile muzzle', 'mobile-eating muzzle', 'mobile-ruminating muzzle' and 'muzzle not visible' are mutually exclusive; the ten ear FAUs are mutually exclusive

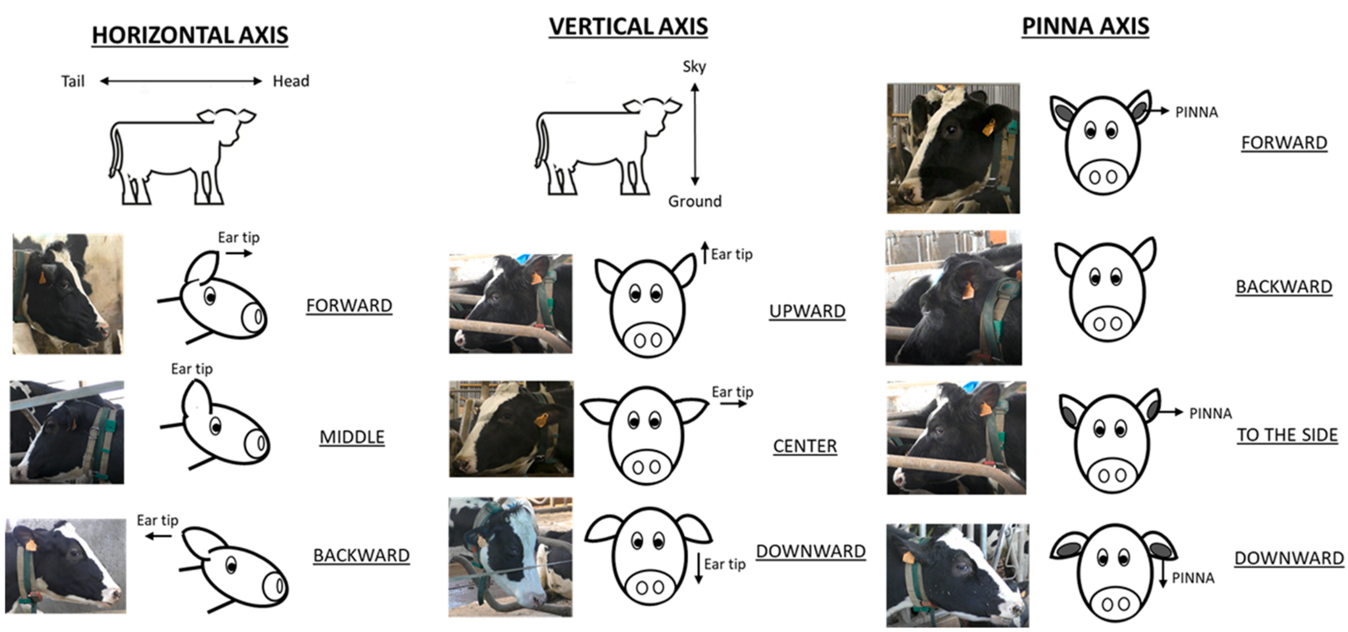
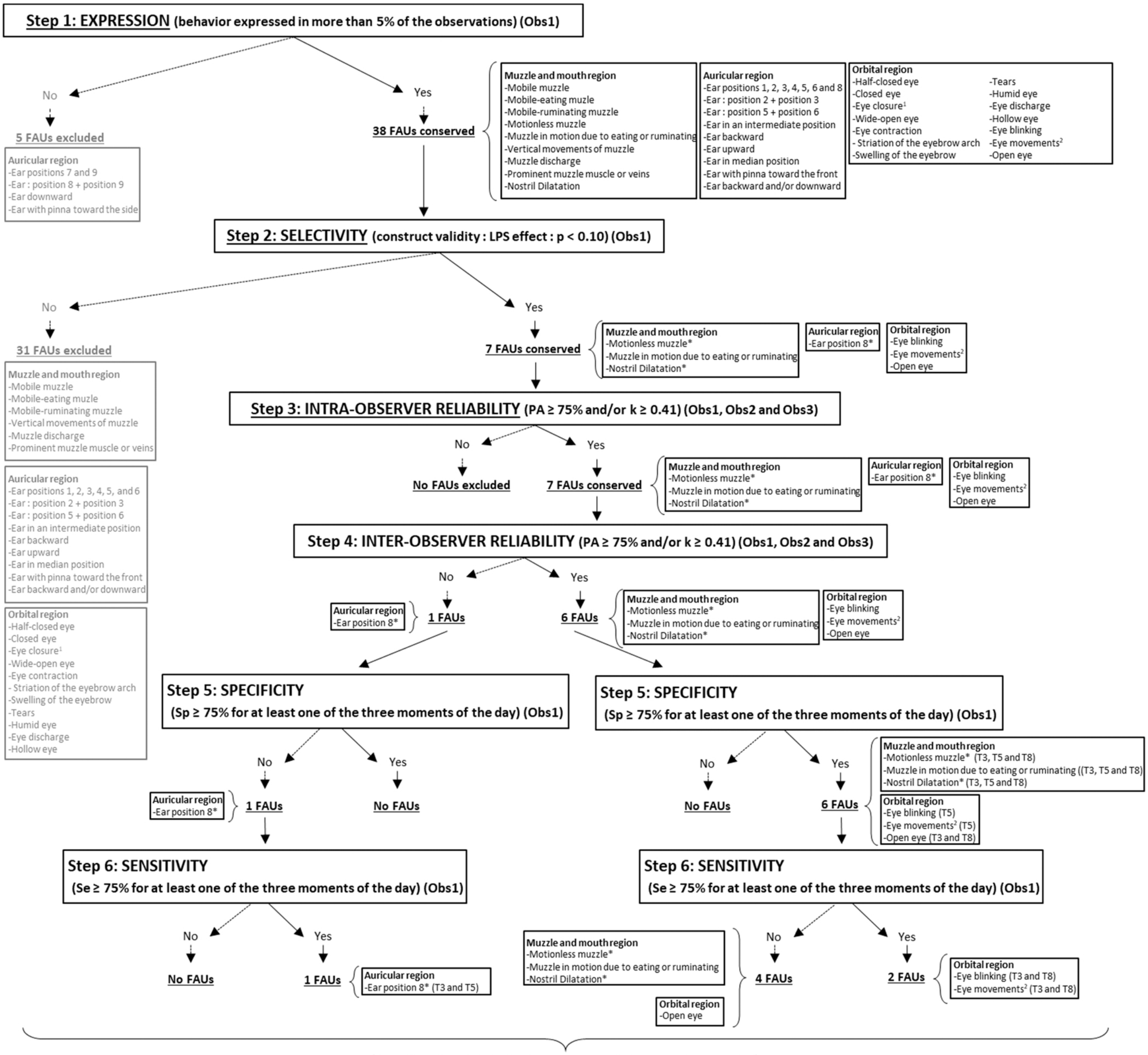


Fig. 2. Photos and illustrations of ear positions according to horizontal, vertical, and pinna axis.

retained for in the next steps. All FAUs that did not differ ($P \geq 0.1$) were considered not selective for detecting pain and thus eliminated from the next steps of the process (Fig. 3).

2.9.2.2. FAU's evaluation after NSAID injection: responsiveness. To investigate the effect of NSAID on FAUs, we used the videos from both treatments (LPS and LPS+NSAID) scored by Obs1. We compared changes in FAUs at the same hour after challenge (i.e. T3 = 11:30 h,



Region	FAUs	Reproductible	Specific	Sensitive
Muzzle and mouth	Motionless muzzle*	Yes	Yes	No
	Muzzle in motion due to eating or ruminating	Yes	Yes	No
	Nostril Dilatation*	Yes	T5 and T8	No
Auricular	Ear position 8*	No	No	T3 and T5
Orbital	Open eye	Yes	T3 and T8	No
	Eye blinking	Yes	T5	T3 and T8
	Eye movements²	Yes	T5	T3 and T8

¹Eye closure = half-closed eye + closed eye
²Eye movements = Eye blinking + Eye contraction

Fig. 3. Six step process to explore the abilities of FAUs to detect pain in dairy cows: video used (videos scored by Obs1/Obs2/Obs3), expression, selectivity, intra-observer reliability, inter-observer reliability, specificity, and sensitivity (Underlined step correspond to eliminative steps; PA: percentage of agreement; k: Fleiss' kappa; * significant effect of NSAID injection with P value ≤ 0.05).

T5 = 13:30 h, and T8 = 16:30 h) between LPS cows vs. LPS+NSAID cows. All the statistical analyses were carried out with nonparametric tests as none of the variables were normally distributed (Normality was checked graphically). Duration or frequency of each FAU after challenge of LPS cows and LPS+NSAID cows were compared using a Mann-Whitney-U test for independent data. Proportion of LPS cows and LPS+NSAID cows assigned to a 'presence' or 'absence' modality after challenge was compared using Fisher tests (Table 2).

2.9.3. *Definition of the pain threshold for each FAU*

To define the pain thresholds for each FAU considered as selective (through construct validity) (see. 2.8.1.1), we used the behavioral data from the videos of LPS cows scored by Obs1. First, the data of duration and frequency were rescaled (by dividing each value by the maximal

Table 2
Statistical methods adapted from COSMIN used to explore the performance characteristics of FAUs in cows exposed to pain: expression, selectivity (construct validity, responsiveness), definition of cut-off, intra-observer reliability, inter-observer reliability, measurement error (specificity, sensitivity).

Aim of the analysis	Definition	Data used for analyses		No of	Statistical Analysis
		Observer	Treatment group	videos used	
Expression	Expression of the FAU	Observer 1	LPS cows and LPS+NSAID cows	156	Descriptive analysis
Selectivity					
Construct validity: Time effect	Alterations in FAUs after LPS infusion	Observer 1	LPS cows	83	Wilcoxon rank-sum test and McNemar tests
Responsiveness: Treatment effect	Alterations in FAUs after ketoprofen injection vs. placebo injection	Observer 1	LPS cows and LPS+NSAID cows	156	Mann-Whitney U test and Fisher tests
Definition of pain threshold	Cut-off value use to transform continuous data into binary variables	Observer 1	LPS cows	83	ROC curves, AUCs and Youden index
Intra-observer reliability	Agreement between coding by the same observer on video scored three times	Each of the 3 observers	LPS cows and LPS+NSAID cows	21	Percentage of agreement and Fleiss' kappa
Inter-observer reliability	Agreement between coding by the three observers	All of the 3 observers	LPS cows and LPS+NSAID cows	156	Percentage of agreement and Fleiss' kappa
Measurement Error					
Specificity	Percentage of LPS cows detected not in pain	Observer 1	LPS cows	83	Number of LPS cows 'detected not in pain' out of the number of LPS cows 'not in pain' (i.e. total number of LPS cows observed before challenge)
Sensitivity	Percentage of LPS cows detected in pain	Observer 1	LPS cows	83	Number of LPS cows 'detected in pain' out of the number of LPS cows 'in pain' (i.e. total number of LPS cows observed after challenge)

value) to obtain continuous data ranging from 0 to 1. All the data point before challenge (T-21 = 11:30 h, T-19 = 13:30 h and T-16 = 16:30 h) were assigned to the negative modality (non painful) and all the data point after challenge (T3 = 11:30 h, T5 = 13:30 h, and T8 = 16:30 h) were assigned to the positive modality (painful). Second, the pain threshold and area under the curve (**AUC**) were determined for each FAU at each time point studied (i.e. T-21 vs. T3 = 11:30 h, T-19 vs. T5 = 13:30 h and T-16 vs. T8 = 16:30 h) using ROC curves and Youden index methods (Swets, 1988; Delacour et al., 2005; Dohoo et al., 2009; Desquilbet, 2022). The Youden index method assigns the same weight to specificity and sensitivity in order to obtain a cut-off with the best possible trade-off between specificity and sensitivity. Therefore, each FAU had three different pain thresholds and AUCs: one per moment of the day (one pain threshold for 11:30 h, a second for 13:30 h and a third for 16:30 h). A ROC curve below the 45° line ruled out defining a Youden index and thus a pain threshold. Third, each observation was allocated either a “detected in pain” or a “detected not in pain” status, defined according to the pain threshold calculated. For instance, if a given behavior increased in painful situation, then a cow displaying that behavior above the pain threshold was considered to be “detected in pain”. Conversely, a cow displaying the behavior below the pain threshold was considered to be “detected not in pain”. Finally, based on these pain thresholds, continuous data (duration for states and frequency for events) were transformed into binary variables (below or above the thresholds).

Table 3 details the AUC, Youden index and pain thresholds for each FAU at each time point (except for 'eye open' and 'nostril dilation' at specific time points, for which it was impossible to define a pain threshold because the ROC curve was below the 45° line) and Fig. S1 the ROC curves for the 8 FAUs at each time point (i.e. T-21 vs. T3 = 11:30 h, T-19 vs. T5 = 13:30 h and T-16 vs. T8 = 16:30 h). For instance, LPS cows spent significantly less time with 'eye open' at T8 after the challenge. At T-16 vs. T8 (16:30 h), the 'eye open' FAU ROC curve that had been drawn had an AUC of 0.65 and a Youden index of 0.09, giving a pain threshold of 36.5 s. Therefore, if cows spent less than 36.5 s (below pain threshold) with 'eye open', they were 'detected in pain' but if they spent more than 36.5 s (above pain threshold) with 'eye open', they were 'detected not in pain'.

2.9.4. *Intra-observer reliability of FAUs: repeatability*

To assess intra-observer reliability, each observer (Obs1, Obs2 and Obs3) scored 21 randomly selected videos (see 2.6 for details). Each observer scored three times each of the 21 videos. Then, in order to obtain binary data, the pain thresholds determined in the previous step (see 2.8.2) at each moment of the day (i.e. thresholds of 11:30 h, 13:30 h and 16:30 h) were applied on continuous data (duration or frequency of the FAUs observed) scored by observers whatever the time point videoed. Therefore, (i) when the three pain thresholds were available, they were applied on each of the 21 videos, creating a total set of 63 observations (21 videos × 3 pain thresholds) per FAU per observer; (ii) when only two pain thresholds were available, both were applied on each of the 21 videos, creating a total set of 42 observations (21 videos × 2 pain thresholds) per FAU per observer; and (iii) when only one pain threshold was available, it was applied on each of the 21 videos, creating a total set of 21 observations (21 videos × 1 pain threshold) per FAU per observer. The statistics for data analysis of intra-observer reliability (repeatability) are detailed in 2.9.5 (see below).

2.9.5. *Inter-observer reliability of FAUs: reproducibility*

To assess the inter-observer reliability of all observers together, we used the final set of 156 videos from LPS cows and LPS+NSAID. Then, in order to obtain binary data, the pain thresholds determined in the previous step (see 2.8.2) at each moment of the day (i.e. thresholds of 11:30 h, 13:30 h and 16:30 h) were applied on continuous data (duration or frequency of FAUs observed) scored by the observers for each corresponding time point (i.e. pain thresholds of 11:30 h were applied on the data from videos at 11:30 h, pain thresholds of 13:30 h were applied on the data from videos at 13:30 h, pain thresholds of 16:30 h

Table 3
Definition of FAU pain thresholds: changes in FAUs by dairy cows in pain, area under curve (AUC), Youden index, pain threshold2 and specificity3 and sensitivity4 estimations for the 14 Holstein dairy cows inoculated with E. coli LPS and injected with saline solution (LPS cows) 30 min after inoculation for all time points combined and for each comparison of time points.

Facial Action Unit	Change of FAUs in pain	Time points studied[5]	No of videos used	AUC	Youden index	Pain threshold	Specificity (%)	Sensitivity (%)
State (in seconds)								
Eye open[6]	Decrease	T-21 vs T3	28	0.59	0.02	39.1	86	36
		T-19 vs T5	28	NC	NC	NC	NC	NC
		T-16 vs T8	27	0.65	0.09	36.5	92	43
Motionless muzzle	Increase	T-21 vs T3	28	0.65	0.09	3.4	93	36
		T-19 vs T5	28	0.68	0.11	4.4	86	50
		T-16 vs T8	27	0.65	0.81	32.2	92	43
Muzzle in motion due to eating or ruminating (Mobile-eating + Mobile-ruminating)	Decrease	T-21 vs T3	28	0.52	1.00	0.0	93	14
		T-19 vs T5	28	0.67	0.12	35.4	79	57
		T-16 vs T8	27	0.52	1.00	0.0	77	43
Ear: position 8 (backward/center/forward)	Decrease	T-21 vs T3	28	0.42	0.56	17.6	14	93
		T-19 vs T5	28	0.66	0.89	4.4	50	86
		T-16 vs T8	27	0.47	0.40	24.1	38	71
Events (in number)								
Eye blinking	Decrease	T-21 vs T3	28	0.46	0.43	8.0	43	79
		T-19 vs T5	28	0.54	1.00	0.0	93	21
		T-16 vs T8	27	0.71	0.65	8.0	46	93
Eye movements (Eye blinking + Eye contraction)	Decrease	T-21 vs T3	28	0.45	0.43	8.0	43	79
		T-19 vs T5	28	0.51	1.00	0.0	93	14
		T-16 vs T8	27	0.68	0.65	8.0	54	93
Nostril dilations	Increases	T-21 vs T3	28	NC	NC	NC	NC	NC
		T-19 vs T5	28	0.61	0.09	1.0	93	29
		T-16 vs T8	27	0.61	0.33	1.0	85	36

[1] Facial action units described in Table 1
[2] In some cases, the Youden index and the threshold could not be calculated = noted "NC" in table
[3] Specificity: number of cows displaying the non-painful modality (based on pain-case hypothesis) of each FAU, out of total number of cows observed at the specific time point
[4] Sensitivity: number of cows displaying the non-painful modality (based on the pain-case hypothesis) of each FAU, out of total number of cows observed at the specific time point
[5] Time points studied: 11:30 h = T-21 vs T3, 13:30 h = T-19 vs T5, 16:30 h = T-16 vs T8
[6] Example for 'Eye open' at T-21 vs T3: in the event of pain, cows decreased the time spent in this FAU, area under curve was 0.57, Youden index was 0.02, pain threshold 39.12 s, Sp at T-21 was 86% (i.e. 86% of cows were 'detected not in pain') and Se at T3 was 38% (i.e. 38% of cows were 'detected in pain').

were applied on the data from videos at 16:30 h).

For each FAU, intra- and inter-observer reliability were assessed on binary data using percentage of agreement (**PA**) and Fleiss' kappa (**k**) coefficient (Table 2). Kappa coefficients are widely used in the literature to assess reliability of categorical data for welfare or pain indicators (e.g. Parham et al., 2019; Lanci et al., 2022). The kappa coefficient allows to take into account agreement due to chance (i.e. it is a chance-adjusted measure of agreement between observers (Byrt et al., 1993). The PA was also calculated as in Parham et al. (2019) by dividing the number of observations for which the observers agreed by the total number of observations and multiplied by 100 (thus, PA = 0 means no agreement and PA = 100 means perfect agreement). The PA was used in addition to Kappa in order to help the interpretation of the Kappa coefficient which can be sometime difficult because the kappa coefficient is affected by the prevalence of the measure and bias between observers (Byrt et al., 1993). Byrt et al. (1993) recommended to not present kappa coefficients alone but with other agreement indicators, as it is now usually done in studies investigating reliabilities of welfare measures (PA being presented along with other reliabilities' coefficients e.g. (Munoz et al., 2018; Vieira et al., 2018; Parham et al., 2019). Percentage of agreement values were interpreted as acceptable if ≥ 75% (Burn et al., 2009; Burn and Weir, 2011; Vieira et al., 2018). Fleiss' kappa coefficients were interpreted as follows: 0.00–0.20 = poor, 0.21–0.40 = passable, 0.41–0.60 = moderate, 0.61–0.80 = substantial, 0.81–1.00 = almost perfect (Landis and Koch, 1977).

FAUs that presented a PA ≥ 75% and/or Fleiss' kappa ≥ 0.41 were considered to have an acceptable reliability (as suggested in Burn et al., 2009; Burn and Weir, 2011; Phythian et al., 2013), and therefore conserved in the next steps. All FAUs that presented a PA < 75% and/or a Fleiss' kappa < 0.41 were not considered repeatable and therefore eliminated from the next step of the process (Fig. 3).

2.9.6. *Specificity and sensitivity of FAUs*

To evaluate specificity (**Sp**, i.e. ability to detect pain-free cows) and sensitivity (**Se**, i.e. ability to detect cows in pain) for each moment of the day (11:30 h i.e. T-21 and T3; 13:30 h i.e. T-19 and T5; 16:30 h i.e. T-16 and T8), we used the 83 videos from LPS cows scored by Obs1. For each FAU, Sp was calculated as the number of cows 'detected as not in pain' out of the number of cows 'not actually in pain' (i.e. total number of LPS cows observed before challenge). FAUs that presented a specificity ≥ 75% for at least one of the three moments of the day (i.e. 11:30 h, 13:30 h and 16:30 h) were considered as sufficiently specific (Fig. 3).

Sensitivity was calculated as the number of cows 'detected in pain' out of the number of cows 'actually in pain' (i.e. total number of LPS cows observed after challenge) (Table 2). For each FAU at each time point, a cow was considered as 'detected in pain' or 'detected not in pain' based on the pain thresholds (determined in 2.8.2), and we considered that cows were actually 'not in pain' at T-21, T-19 and T-16 but 'in pain' at T3, T5 and T8. FAUs that presented a sensitivity ≥ 75% for at least one of the three moments of the day (i.e. 11:30 h, 13:30 h and 16:30 h) were considered as sufficiently sensitive (Fig. 3).

All analyses were performed using R software version 3.6.2 (2019). The *ROCR* package (Sing et al., 2005) was used to plot ROC curves, the *AUC* package was used to estimate AUCs, the *irr* package was used to estimate PA, and the *DesTools* package was used to estimate Fleiss' kappa coefficient and confidence interval (**CI**).

For McNemar, Fisher and non-parametric tests, a P-value of < 0.05 was considered as statistically significant whereas P-values of 0.05 ≤ P < 0.10 were considered a trend.

3. Results

Fig. 3 presents the six-step process followed for the 43 FAUs. Results at each step are detailed below.

3.1. Expression of FAUs by cows

Among all FAUS, two were expressed by the cows in less than 5% of the videos: 'ear position 7′ and 'ear position 9′. 'Ear position 7′ was observed by Obs 1 in five videos and 'Ear position 9′ was observed by Obs 1 in only two videos among 156. These two FAUs were thus excluded for the next steps (Fig. 3). Moreover, 3 FAUs which resulted from the combination based on ear positions 9 and another FAU were also removed: 'ear position 8 + 9′; 'Ear downwards' which corresponded to ear positions 4 + 9; and 'Ear with pinna toward the side' which corresponded to ear positions 6 + 9. Therefore, 5 FAUs over 43 were removed after the first step of the process (Fig. 3).

3.2. Selectivity of FAUs for assessing pain

3.2.1. FAUs changes in response to E.coli LPS infusion in the udder (construct validity)

Changes in FAUs in LPS cows after LPS infusion are detailed in Table S2 and Table S3.

Regarding the orbital region, compared to before LPS infusion (T-16 = 16:30 h), LPS cows at T8 = 16:30 h spent significantly less time with 'eye open' (T-16: 40.0 [13.6–40.0] seconds; T8: 40.0 [17.2–40.0] seconds, P = 0.036): the median of the duration of 'eyes open' did not differ between T-16 and T8, however at T-16 there was no variance because all cows had the eyes open for all the 40 ss, and at T8 this was not the case (IR: 17.2–40.0). Compared to before LPS infusion (T-16), LPS cows at T8 also tended to display less 'eye blinking' (T-16: 7.0 [4.0–11.0]; T8: 3.0 [2.0–6.0], P = 0.071) and less 'eye movements' (T-16: 10.0 [4.0–11.0]; T8: 4.0 [2.0–6.0], P = 0.091).

Regarding the muzzle and mouth region, compared to before LPS infusion (T-19 = 13:30 h), LPS cows at T5 = 13:30 h spent significantly less time with 'muzzle in motion due to eating or ruminating' (T-19: 36.2 [35.8–39.4] seconds; T5: 35.3 [0.00–37.9] seconds, P = 0.045), and they tended to spend more time with 'motionless muzzle' (T-19: 3.48 [0.00–4.10] seconds; T5: 4.28 [3.32–37.3] seconds, P = 0.068). Compared to before LPS infusion (T-19 = 13:30 h), LPS cows at T5 = 13:30 h also tended to display more 'nostril dilation' (T-19: 0.00 [0.00–0.00]; T5: 0.00 [0.00–0.75], P = 0.097); the median of the frequency of 'nostril dilation' did not differ between T-19 and T5, however at T-19 there was no variance because no cow displayed that behavior, while they did at T5 (IR: 0.00–1.0).

Regarding the auricular region, compared to LPS infusion (T-19 = 13:30 h), LPS cows at T5 = 13:30 h tended to spend less time in 'ear: position 8 (backward/center/forward)'(T-19: 10.3 [0.00–37.4] seconds; T5: 0.00 [0.00–0.00] seconds, P = 0.057.).

Out of the 43 FAUs tested for selectivity (through construct validity), 7 (i.e. 'eye open', 'eye blinking', 'eye movements', 'motionless muzzle', 'muzzle in motion', 'nostril dilation', and 'ear: position 8′) were considered as selective at least one time point, and therefore conserved for the next steps. All the other FAUs were eliminated (Fig. 3).

3.2.2. FAUs changes in LPS+NSAID cows compared to LPS cows (responsiveness)

The post-challenge changes in LPS+NSAID cows' FAUs compared to LPS cows' FAUs are detailed in Table S4 and Table S5.

Regarding the orbital region, LPS+NSAID cows tended to spend more time with 'eye open' at T3 = 11:30 h and T8 = 15:30 h (respectively P = 0.089 and P = 0.067) than LPS cows.

Regarding the muzzle and mouth region, LPS+NSAID cows spent significantly less time with 'motionless muzzle' at T5 = 13:30 h and T8 = 15:30 h (respectively P = 0.023 and P = 0.028) than LPS cows. LPS+NSAID cows displayed significantly less 'nostril dilation' (P = 0.022) at T8 than LPS cows. Moreover, LPS+NSAID cows also tended to spend more time with 'muzzle in motion' (P = 0.071) at T5 than LPS cows.

Regarding the auricular region, LPS+NSAID cows spent significantly more time in 'ear: position 8 (backward/center/forward)' at T5 = 13:30 h (P = 0.011).

3.3. Intra-observer reliability of FAUs: repeatability

Table 4 details the **PA** and **k** [95% CI] values for intra-observer reliability (Table S6).

The three observers reached acceptable PA and/or acceptable k for all FAUS's (i.e. 'eye open', 'eye blinking', 'eye movements', 'motionless muzzle', 'muzzle in motion', 'nostril dilation', and 'ear: position 8′ (backward/center/forward)) (Table 4).

All 7 of the FAUs tested for intra-observer reliability (i.e. 'eye open', 'eye blinking', 'eye movements', 'motionless muzzle', 'muzzle in motion', 'nostril dilation', and 'ear: position 8′) were thus considered as repeatable and therefore retained for the next steps (Fig. 3).

3.4. Inter-observer reliability of FAUs: reproducibility

Table 4 details the PA and k [95% CI] values for inter-observer reliability (Table S7).

Regarding the orbital region, the FAUs 'eye blinking' and 'eye movements' had acceptable inter-observer reliability. The FAU 'eye open' had insufficient PA (72%) but acceptable k (0.51) (Table 4).

Regarding the muzzle and mouth region, the FAUs 'muzzle in motion', 'motionless muzzle' and 'nostril dilation' had acceptable PA and acceptable k (Table 4).

Regarding the auricular region, the FAU 'ear: position 8 (backward/center/forward)' had insufficient PA (64%) and insufficient k (0.26) (Table 4).

3.5. Specificity and sensitivity of the FAUs

Table 3 details Sp and Se for each FAU.

Regarding the orbital region, the 'eye open' FAU was specific at T3 (11:30 h) and T8 (16:30 h) (Sp ≥ 86%) but was not sensitive (Se ≤ 43%). The 'eye blinking' and 'eye movements' FAUs were specific at T5 (13:30 h) (Sp = 93%) and sensitive at T3 (11:30 h, Se = 79%) and T8 (16:30 h, Se = 93%).

Regarding the muzzle and mouth region, three FAUs (i.e. 'motionless muzzle', 'muzzle in motion due to eating or ruminating' and 'nostril dilation') were specific (Sp ≥ 77%) for all time points, but were not sensitive (Se ≤ 57%).

Regarding the auricular region, the FAU 'ear: position 8 (backward/center/forward)' were not specific (Sp ≤ 50%) at all time points, and were sensitive (Se ≥ 86%) at T3(11:30 h) and T5 (13:30 h) but not at T8 (=16:30, Se = 71%).

Among the 7 FAUs tested for specificity, 6 FAUs ('eye open', 'eye blinking', 'eye movements', 'motionless muzzle', 'muzzle in motion', 'nostril dilation') were considered sufficiently specific for at least one time point, while 'ear: position 8′ was considered not sufficiently specific for all time point (Fig. 3).

Among the 7 FAUs tested for sensitivity, 3 ('eye blinking', 'eye movements' and 'ear: position 8′) were considered sufficiently specific for at least one time point, while 'eye open', 'motionless muzzle' 'muzzle in motion due to eating or ruminating', and 'nostril dilation' were considered not sufficiently sensitive for all time points (Fig. 3).

At the end of the process, 7 FAUs (3 orbital-region FAUs, 3 muzzle and mouth-region FAUs, and 1 auricular-region FAUs) had partially satisfied all the performance-characteristics (Fig. 3).

Table 4
Intra-observer reliability and inter-observer reliability of each FAUs1, in 27 Holstein-Friesian dairy cows inoculated with E. coli LPS, half of which were injected with saline solution (LPS cows, n = 14) and the other half with 3 mg/kg BW ketoprofen (LPS+NSAID cows, n = 13) half an hour after inoculation. Absolute agreement and Fleiss' kappa were calculated to test if each observer coded identically between replicated videos and if the three observers coded identically for each 40-second video.

	Intra-observer reliability							Inter-observer reliability		
		Percentage of agreement (in %)			Fleiss' kappa [95% CI]				Percentage of agreement (in %)	Fleiss' kappa [95% CI]
Facial Action Unit	n[2]	Median	Min	Max	Median	Min	Max	n[3]		
State (in seconds)										
Eye open[3]	42	95.2	69.0	95.2	0.94 [0.83–1.00]	0.53 [0.35–0.71]	0.94 [0.83–1.00]	103	71.8	0.51 [0.37–0.65]
Motionless muzzle	63	87.3	87.3	95.2	0.82 [0.70–0.93]	0.82 [0.70–0.93]	0.93 [0.85–1.00]	156	87.2	0.73 [0.61–0.83]
Muzzle in motion due to eating or ruminating (Mobile-eating + Mobile-ruminating)	42	95.2	92.9	95.2	0.93 [0.82–1.00]	0.90 [0.77–1.00]	0.93 [0.82–1.00]	156	94.9	0.90 [0.82–0.96]
Ear: position 8 (backward/center/forward)	63	87.3	82.5	98.4	0.77 [0.65–0.87]	0.69 [0.46–0.85]	0.95 [0.84–1.00]	156	64.1	0.26 [0.15–0.38]
Events (in number)										
Eye blinking	42	95.2	92.9	100	0.94 [0.83–1.00]	0.90 [0.78–1.00]	1.00 [1.00–1.00]	156	74.4	0.66 [0.57–0.75]
Eye movements (Eye blinking + Eye contraction)	42	97.6	92.9	100	0.97 [0.88–1.00]	0.90 [0.78–1.00]	1.00 [1.00–1.00]	156	77.6	0.70 [0.62–0.79]
Nostril dilations	21	100	90.5	100	1.00 [1.00–1.00]	0.58 [0.48–0.98]	1.00 [1.00–1.00]	105	93.3	0.54 [0.36–0.79]

[1] Facial action units described in Table 1
[2] n = number of videos coded by each observer
[3] Example for 'Eyes open': for intra-observer reliability, each observer coded 21 videos on which we applied two pain thresholds (11:30 h corresponding to T-21 and T3, and 16:30 h corresponding to T-16 and T8), leading to 42 units of observations. For inter-observer reliability, each observer coded 103 s videos

4. Discussion

4.1. Key messages

The use of FAUs changes to detect pain have been explored in many species, but their performance-characteristics have been under-investigated. Here we bring a six-step process of behavioral indicators of pain adapted from COSMIN methodology (Mokkink et al., 2016, 2018; Tomacheuski et al., 2023), including (1) expression, (2) selectivity (through construct validity and responsiveness), (3) intra-observer reliability, (4) inter-observer reliability, (5) specificity and (6) sensitivity. This process was applied to 43 FAUs in primiparous dairy cows with LPS-induced udder inflammation (Ginger et al., 2023). The design of the experiment allowed us to take into account the "individual" effect, by using each cow as her own control; but also the "circadian" effect, by comparing cows responses at the same moment of the day before vs after challenge. This study is the first, to our knowledge to explore FAUs at different moments of the day. Over the 43 FAUs tested, seven FAUs partially satisfied the process. More precisely, six FAU were expressed, selective (through construct validity), repeatable, reproducible, and either specific or sensitive (Fig. 3): three from the orbital region ('eye blinking', 'eye movements', 'eye open'), and three from the muzzle-and-mouth region ('nostril dilation', 'motionless muzzle', 'muzzle in motion due to eating/ruminating'). One FAU from the auricular region ('ear: position 8′, i.e. ears backwards on the cranio-caudal axis, in the center of the dorso-ventral axis, and the pinna is to the side) was expressed, selective (through construct validity), repeatable and sensitive but not reproducible nor specific. This study proposes a list of potential FAU that could be included in a pain grimace scale in cattle. Further studies are warranted to validate the FAUs alone or combined, in other pain models and with a larger set of cows.

4.2. A six step process applied to FAUs to detect pain in cattle

We choose to test (1) expression, (2) selectivity (through construct validity and responsiveness), (3) intra-observer reliability, (4) inter-observer reliability, (5) specificity and (6) sensitivity as a selection process using a waterfall method. Expression, selectivity and intra-observer reliability were eliminative steps as we considered these performance-characteristics to be the most crucial within the process. Indeed, if an indicator is too rare, or does not measure what it is supposed to measure, it is not relevant. Moreover, if an observer is not able to obtain the same results on the same observations the measure is not reliable at all. The inter-observer reliability was not an eliminative step as we considered that a poor inter-observer reliability could result from insufficient training and, could probably be improved thanks to additional training. The specificity and the sensitivity were not eliminative steps as we considered that the decision about the acceptability of Se and Sp rely on the context in which indicators are used.

4.3. FAUs that were not expressed

Over the 43 FAUs investigated, two were excluded because they were too rare, i.e. they were expressed in less than 5% of the videos: ear position 9 (backward / downward / pinna to the side) and ear position 7 (backward / upward / pinna downward)). Our results confirm that cows seem to rarely display 'backward / downward / pinna to the side' ear position (i.e. ear position 9) as already shown in cows experiencing either neutral, positive (joy) or negative (frustration) emotional state (i.e. ear position 4 in Lambert and Carder, 2019). Ear position 7, was rarely displayed here even before challenge while it was described (i.e. ear position 1 in Lambert and Carder, 2019) in cows before feeding, but also while experiencing joy/excitement during food consumption (Lambert and Carder, 2019). The context may explain this difference: in Lambert and Carder (2019)'s study, the cows were performing a behavioral test with different food thus eliciting high arousal positive/negative emotional state, while in the present study cows were left in their home pen, experiencing certainly a lower arousal emotional state, either neutral (before challenge) or negative (after challenge). Further studies are therefore needed to explore the association between ear postures and arousal.

4.4. FAUs that lacked selectivity for assessing pain associated with udder inflammation

Over the 41 FAUs investigated, 35 were excluded because of a lack of selectivity (through construct validity). After the challenge, cows did not close their eyes more nor open them wider, did not present 'swelling of the eyebrow', 'hollow eye', 'striations of the eyebrow arch', 'muzzle discharge' and showed no change in 'ear backward' and 'ear downward'. These results were surprising, as cattle in pain were reported to reduce their orbital tightening (Yamada et al., 2021), to have hollow eyes with furrow lines above (Gleerup et al., 2015a), a tension of the orbital region or pronounced bony relief above the eyes (Müller et al., 2019; Yamada et al., 2021), prominent strained chewing muscles (Gleerup et al., 2015a; Yamada et al., 2021), their ears backward and/or downward (Gleerup et al., 2015a; Müller et al., 2019). Here, the lack of changes in some FAUs following the challenge could be explained by the pain model, the experimental design, the coding support, the coding method, and/or the cows' face coloring.

Previous studies used painful models inducing more intense (e.g. castration (Yamada et al., 2021)), and long-lasting (e.g. lameness or respiratory disease (Gleerup et al., 2015a)) pain. Here, the cows suffered from a mild mastitis with an acute phase lasting 8 h (Ginger at al., 2023). To our knowledge this study is one of the first to show that some FAUs may be used to identify mild pain in cattle, while in horses, mild inflammation was not able to be detected by facial indicators, despite clinical and pathophysiological signs (Carvalho et al., 2022). Therefore, changes in certain FAUs may depend on intensity of the pain model (mild vs. severe) per se.

Our experimental design, with cows being their own control, allowed us to explore the cows' individual FAUs changes to the noxious stimulus, whereas in Gleerup et al. (2015a) and Yamada et al. (2021) animals from different groups were compared. Therefore, changes in certain FAUs may vary according to individual variability.

Previous studies used still images (Müller et al., 2019; Yamada et al., 2021). Video coding, as used here, allows coding of FAUs durations and frequencies of very-short-lived facial expressions. Nevertheless, coding a moving animal may be more challenging than coding a still image, explaining why selectivity here diverge from previous findings. In addition, the videos used in the present study were taken at precise time points without selecting a priori behavioral responses while in most studies images were selected by researchers from videos (e.g. Yamada et al., 2021), probably enhancing potential selection bias (through the selection of images with most intense facial expression).

Our study is the first to investigate quantitative changes of FAUs. Previous research used two methods of coding FAUs: the points-score scale (e.g. McLennan et al., 2016), and presence/absence of the item (Gleerup et al., 2015a; Müller et al., 2019). Differences in observational methods could lead to inconsistent findings between studies.

Finally, most of our cows had large black patches on their head, unlike Nellore bulls (Yamada et al., 2021). Like in horses (Dalla Costa et al., 2016) and in cats (Evangelista et al., 2019), black animals are more difficult to observe than white ones, especially on orbital and muzzle regions. Therefore, some FAUs were coded 'not visible' especially when light was low.

4.5. FAUs that were selective, repeatable, and reproducible

Six FAUs were selective (through construct validity), repeatable, and reproducible: 'eye open', 'eye blinking', 'eye movements', 'motionless muzzle', 'muzzle in motion due to eating or ruminating' and 'nostril dilatation'. They were selective from 5 h (T5) to 8 h (T8) after LPS infusion and were concomitant with pathophysiological and behavioral modifications (Ginger et al., 2023).

Regarding the orbital region, at T8, frequencies of 'eye blinking' and 'eye movements' decreased in LPS cows. No change in 'eye blinking' frequency were reported in horses exposed to two noxious stimuli on forelimbs (Gleerup et al., 2015b). After challenge, LPS cows spent less time with 'eye open'. There was no significant difference between LPS vs LPS+NSAID cows. However, the challenge had no effect on time spent with 'half-closed eye', 'closed eye' or 'wide-open eye' in LPS cows. These results are not congruent with the presence of orbital tightening already observed in pain contexts in horse (Dalla Costa et al., 2014) and cattle (Gleerup et al., 2015a; Yamada et al., 2021). These differences could be explained by the observational methods used (point-score scale vs duration) and/or the coding support (image vs videos). Our study is the first to detail the time spent in the different stages of eye opening over 40-second observations, whereas most studies have only reported the absence/moderate/obvious presence of orbital tightening (e.g. Dalla Costa et al., 2014; McLennan et al., 2016; Yamada et al., 2021).

Regarding muzzle, at T5, LPS cows spent significantly less time with 'muzzle in motion' and tended to spend more time with 'motionless muzzle'. Moreover, both changes were counteracted by NSAID injection, thus confirming their high selectivity in the context of pain. Following the challenge, LPS cows tended to display more dilated nostril events. Compared to LPS cows, LPS+NSAID cows displayed significantly less 'nostril dilation'. Therefore 'nostril dilation' could be selective for pain detection in cows, in line with previous results in horses (Dalla Costa et al., 2014; Lanci et al., 2022) and cattle (Müller et al., 2019; Yamada et al., 2021). Further research could include this FAU to detect pain in animals.

4.5.1. FAUs that were also specific

Six FAUs were specific at particular time points: 'eye open' (at T3, T8), 'eye blinking' (at T5), 'eye movements' (at T5), 'motionless muzzle' (at T3, T5, T8), 'muzzle in motion due to eating or ruminating' (at T3, T5, T8) and 'nostril dilatation' (at T5, T8). Based on their good selectivity, intra- and inter-observer reliability and specificity $> 75\%$ at particular time points, these six FAUs could be considered 'acceptable' FAUs to be used in a grimace scale. Nevertheless, their Se was $\leq 57\%$ at the corresponding time points, which could be explained by the mild pain experienced by the cows (Ginger et al., 2023). This low Se could also come from the method used to establish the cut-off, in order to transform continuous data into binary data. This method assigns the same weight to Sp and Se in order to obtain a cut-off with the best possible trade-off (Desquilbet, 2022). Depending on the purpose of facial expression observation, greater weight could be given to either Se or Sp. For instance, if the aim is to detect any animal in pain, then more importance should be given to Se. These selective, repeatable, and reproducible and specific FAUs would therefore be good candidates for pain grimace scale in cows because they allow detecting pain free animals, but their ability to identify animals in pain is insufficient. Further studies should explore their combination with indicators with high sensitivity in a grimace scale.

4.5.2. FAUs that were also sensitive

Among the 43 FAUs studied, two FAUs were selective (through construct validity), repeatable, reproducible and sensitive at particular time points: 'eye blinking' (at T3 and T8) and 'eye movement' (T3, T8). Based on their good selectivity, intra and inter-observer reliability and sensitivity $> 75\%$ at specific time points, 'eye blinking' and 'eye movements' FAUs could be considered 'acceptable' to be used in a grimace scale. Nevertheless, their Sp was $\leq 54\%$ at the corresponding time points. These FAUs would be good candidates for pain grimace scale in cows because they allow detecting painful animals, even if their ability to identify animals free of pain is insufficient. Further studies should explore their combination with indicators with high specificity in a grimace scale.

4.6. FAUs that were selective, repeatable, sensitive, but not reproducible

Among the 43 FAUs studied, one ('ear: position 8′, i.e. the ear is backwards on the cranio-caudal axis, in the center of the dorso-ventral axis, and the pinna is to the side)) was selective (through construct validity), repeatable, specific, and sensitive, but not reproducible. The LPS cows tended to spend less time in this ear position after challenge, contrary to previous research in which animals experiencing pain had ears backward (e.g. Dalla Costa et al., 2014; Müller et al., 2019).

Their inter-observer reliability was low, unlike high inter-observer reliability of 'stiffly backwards ears' in horses (Dalla Costa et al., 2014). This could come from the horses been submitted to castration, and their FAUs were scored on images using a 3-points scale. The poor reproducibility of these two ear positions could also be related to over-precision of the ethogram that contained ten ear FAUs which could make the observations over-complex. Moreover, rather than scoring two or three positions on a still image, observers had to score them simultaneously on a video. Further research with a simpler ethogram is therefore needed.

4.7. Can we use FAUs at any time of the day?

The performances of FAU depended upon the time of the day: FAUs were either specific or sensitive, but not both simultaneously. More specifically, i) 'ear position 8′', was sensitive but not specific at T3 = 11:30 h and T5 = 13:30 h; ii) 'ear position 8′ was neither sensitive nor specific at T8; iii) 'eye open', 'motionless muzzle', 'motion in motion due to eating or ruminating', 'nostril dilatation' were specific but not sensitive whatever the time; iv) 'eye blinking', 'eye movement' were specific but not sensitive at T5 = 13:30 h, while they were sensitive but not specific at T3 = 11:30 h and T8 = 16:30 h. To our knowledge this study is the first to highlight measurement error (sensitivity and specificity) according to the moment at which the cows are observed: no FAU were both specific and sensitive at the same time. Our results show that a given FAU, according to the hour of the day, could be either specific or sensitive, but not specific and sensitive at the same time. This might be explained either by FAUs changes being concomitant with the acute phase of inflammation (T3 to T8) (Ginger et al., 2023) or because these FAUs could depend upon the nycthemeral rhythm of cows and therefore only modified at certain moment of day. Moreover, all AUC values were close to 0.5 (ranging from 0.42 to 0.71), indicating that their discrimination between painful vs pain-free cows was quite low (Swets, 1988). Our results therefore demonstrate that no FAU satisfied all performances characteristic. This highlights the necessity to combine several indicators (here FAUs); as already shown in other contexts (Faure et al., 2017) and/or in other species using FAU (Dalla Costa et al., 2018) to efficiently detect pain in animals.

4.8. Study limitations

The present study presents several limitations.

First, we used a *E. coli* O111 LPS mammary challenge that produced mild clinical, pathophysiological and behavioral responses (Ginger et al. 2023). This design allowed us to associate some FAU with mild pain which has - to our knowledge - not been performed in other species (e.g. horses, Carvalho et al., 2022) as most grimace scales were developed under moderate or severe pain (Mogil et al., 2020; Evangelista et al., 2022). However, the design of the present study, based only on mild mastitis, could not allow exploring whether the FAU used here would also be relevant in other pain levels or in other pain contexts (e.g. mild lameness).

Second, to ensure the best quality video possible, especially in cubicle barns with multiple rails and partitions, recording was performed by a human using a camcorder, which implied that the person had to stand still at the boundary of the home pen. Although pain related facial expressions cannot be completely suppressed by voluntary control (e.g. in humans, (Williams, 2002)), studies on horses (Torcivia and McDonnell, 2020) and rabbits (Pinho et al., 2020, 2023) have demonstrated that in the presence of a human the animals displayed less in facial expression of pain or discomfort behavior than when the human was not present. Therefore, the human presence in our study may have prevented cows from displaying facial expression of pain.

Third, we used the data from one observer (Obs1) to check FAU's expression and selectivity (through construct validity and responsiveness), and to evaluate Specificity and Sensitivity. Data from Obs1, Obs 2 and Obs 3 were used only to test intra-observer reliability (repeatability) and inter-observer reliability (reproducibility). This may have decreased the power of the process. However, we had chosen the most experienced observer as a reference who had high agreement with the two other observers.

Fourth, this study was based on video edition and analysis. The videos used lasted 40 s, and 32 items were coded. This procedure allowed detailed analysis of FAU's, but prevented its direct applicability on the field. Other studies used photographs extracted from videos lasting longer, i.e 3 min (Evangelista et al., 2019), 20 min (Gleerup et al., 2015a), 30 min (Dalla Costa et al., 2014). Further work exploring the simplification of the procedure used here is needed, to provide an on-field tool for farmers, veterinarians and technicians. Such study could for instance compare the outcomes of the present study to ones from the same 40 s videos scored using a presence/absence grid.

Fifth, among the three observers, two were female and one was male. Ideally, we should have balanced between gender with two men and women. Nevertheless, inter-observer reliability was high for most of the FAUs explored here. Gender seems indeed to influence pain evaluation: for instance women attributed higher pain scores than men when assessing pain in cats after surgery (Luna et al., 2022) or with various medical conditions (Evangelista and Steagall, 2021). However, these studies compared total pain score attributed by gender but the gender comparison for each FAU were not provided. Further studies with more observers should explore the effect of gender on cattle pain assessment.

4.9. Perspectives

In the present study we did not explore feasibility of FAUs. This could be included as a step of validation in further studies, by e.g. comparing feasibility on cows of different colors, on different kind of farm housing. The present study used videos and precise thresholds in duration and frequency, which certainly precludes its feasibility on the field. Thus, our FAUs could be used in an experimental context using video recording, but not in "live" observation. The next step would compare the results between detailed video observation (as done here), and a scoring of each FAU after visioning 40 s videos, as previously done in cats (Evangelista et al., 2020) and rodents (Miller and Leach, 2015; Leung et al., 2016). For instance, observers would be asked to record if cows displayed or not muzzle movement during the video, instead of its precise duration. Once the FAUs validated for feasibility they would be included in a composite grimace scale. Future studies would then be needed to complete the validation process, by performing: internal consistency (principal component analysis or factor analysis, item-total correlation, Cronbach's alpha coefficient, and McDonald's omega coefficient), criterion validity (correlation with other instruments), weightings of items, the optimal cutoff point and AUC of the complete instrument, and diagnostic uncertainty zone.

5. Conclusion

Among 43 FAUs explored individually, 7 FAUs from the three regions of the head (orbital, muzzle-and-mouth, and auricular) were identified as potential candidates for detecting cows in pain. Four of these 7 FAUs ('eye open', 'motionless muzzle', 'muzzle in motion' and 'nostril dilatation') satisfied all following steps of the process except sensitivity. Two other FAUs ('eye blinking', 'eye movements') satisfied selectivity, intra- and inter-observer reliability criteria and, depending on the time point considered, were either sensitive or specific but not both at the same time. The last FAU ('ear: position 8′) satisfied selectivity, intra-observer reliability and sensitivity criteria but not inter-observer reliability and specificity criteria. Future studies are now needed to complete this process with the other validation steps to assess the set of FAUs into an instrument. The performances could be also explored on longer periods and in several other pain contexts. Also, further studies are needed to explore the use of such FAUs during veterinary examination on commercial farms, or in veterinary hospitals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is part of the LongHealth project that received financial support from INRAE (GISA [Gestion Intégrée de la Santé Animale] metaprogram), and from Conseil Régional Auvergne-Rhône-Alpes (CPER 2015-2020 PhD grant). The NSAID used was kindly supplied by CEVA Santé Animale. The authors thank Charlotte Bagnard for her help with video-recording, Eric Delval and Cecilia Brooks for coding the videos, Romain Lardy for his precious help with pain threshold definition, Marie-Laure Delignette-Muller for her advices in statistics, Isabelle Rautenbach, Gilles Foucras and Pierre Germon for their help throughout the project, and Caroline Boulocher and Didier Pin for their help with cattle facial anatomy. We also thank Metaform for valuable English-language proofing support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2023.105951.

References

Banchi, P., Quaranta, G., Ricci, A., von Degerfeld, M.M., 2020. Reliability and construct validity of a composite pain scale for rabbit (CANCRS) in a clinical environment. PLoS One 15, e0221377. https://doi.org/10.1371/journal.pone.0221377.

de Boyer Des Roches, A., Faure, M., Lussert, A., Herry, V., Rainard, P., Durand, D., Foucras, G., 2017. Behavioral and patho-physiological response as possible sings of pain in dairy cows during Escherichia coli mastitis: a pilot study. J. Dairy Sci. 100, 8385–8397.

Burn, C.C., Weir, A.A.S., 2011. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. Vet. J. 188, 166–170. https://doi.org/10.1016/j.tvjl.2010.04.021.

Burn, C.C., Pritchard, J.C., Whay, H.R., 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Anim. Welf. 18, 177–187. https://doi.org/10.1017/S0962728600000324.

Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46, 423–429. https://doi.org/10.1016/0895-4356(93)90018-v.

Carvalho, J.R.G., Trindade, P.H.E., Conde, G., Antonioli, M.L., Funnicelli, M.I.G., Dias, P. P., Canola, P.A., Chinelatto, M.A., Ferraz, G.C., 2022. Facial expressions of horses using weighted multivariate statistics for assessment of subtle local pain induced by polylactide-based polymers implanted subcutaneously. Animals 12, 2400. https://doi.org/10.3390/ani12182400.

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., Leach, M.C., 2014. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS One 9, e92281. https://doi.org/10.1371/journal.pone.0092281.

Dalla Costa, E., Pascuzzo, R., Leach, M.C., Dai, F., Lebelt, D., Vantini, S., Minero, M., 2018. Can grimace scales estimate the pain status in horses and mice? A statistical approach to identify a classifier. PLoS One 13. https://doi.org/10.1371/journal.pone.0200339.

Delacour, H., Servonnet, A., Perrot, A., Vigezzi, J.F., Ramirez, J.M., 2005. La courbe ROC (receiver operating characteristic): principes et principales applications en biologie clinique. Ann. Biol. Clin. 63.

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 - Altern. Anim. Exp. 34, 409–429. https://doi.org/10.14573/altex.1607161.

Desquilbet, L.L. 2022. Tutoriel sur les courbes ROC et leur création grâce au site Internet easyROC. doi:HAL Id: hal-02870055.

Di Giminiani, P., Brierley, V.L.M.H., Scollo, A., Gottardo, F., Malcolm, E.M., Edwards, S. A., Leach, M.C., 2016. The assessment of facial expressions in piglets undergoing tail docking and castration: toward the development of the piglet grimace scale. Front Vet. Sci. 3. https://doi.org/10.3389/fvets.2016.00100.

Dohoo, I., W. Martin, and H. Stryhn. 2009. Veterinary Epidemiologic Research, 2nd Edition. Charlottetown, Canada.

Evangelista, M.C., Steagall, P.V., 2021. Agreement and reliability of the Feline Grimace Scale among cat owners, veterinarians, veterinary students and nurses. Sci. Rep. 11, 5262. https://doi.org/10.1038/s41598-021-84696-7.

Evangelista, M.C., Monteiro, B.P., Steagall, P.V., 2022. Measurement properties of grimace scales for pain assessment in nonhuman mammals: a systematic review. PAIN 163, e697. https://doi.org/10.1097/j.pain.0000000000002474.

Evangelista, M.C., Watanabe, R., Leung, V.S.Y., Monteiro, B.P., O'Toole, E., Pang, D.S.J., Steagall, P.V., 2019. Facial expressions of pain in cats: the development and validation of a Feline Grimace Scale. Sci. Rep. 9, 19128. https://doi.org/10.1038/s41598-019-55693-8.

Evangelista, M.C., Benito, J., Monteiro, B.P., Watanabe, R., Doodnaught, G.M., Pang, D.S. J., Steagall, P.V., 2020. Clinical applicability of the Feline Grimace Scale: real-time versus image scoring and the influence of sedation and surgery. PeerJ 8, e8967. https://doi.org/10.7717/peerj.8967.

Faure, M., Paulmier, V., Boissy, A., de la Foye, A., de Boyer Des Roches, A., Durand, D., 2017. A multiparametric approach to discriminate the impacts of different levels of invasiveness of surgical procedures in sheep. Animal 1–10.

Fitzpatrick, C.E., Chapinal, N., Petersson-Wolfe, C.S., DeVries, T.J., Kelton, D.F., Duffield, T.F., Leslie, K.E., 2013. The effect of meloxicam on pain sensitivity, rumination time, and clinical signs in dairy cows with endotoxin-induced clinical mastitis. J. Dairy Sci. 96, 2847–2856. https://doi.org/10.3168/jds.2012-5855.

Ginger, L., Ledoux, D., Bouchon, M., Rautenbach, I., Bagnard, C., Lurier, T., Foucras, G., Germon, P., Durand, D., de Boyer des Roches, A., 2023. Using behavioral observations in freestalls and at milking to improve pain detection in dairy cows after lipopolysaccharide-induced clinical mastitis. J. Dairy Sci. https://doi.org/10.3168/jds.2022-22533.

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

Gleerup, K.B., Forkman, B., Lindegaard, C., Andersen, P.H., 2015b. An equine pain face. Vet. Anaesth. Analg. 42, 103–114. https://doi.org/10.1111/vaa.12212.

Guesgen, M.J., Beausoleil, N.J., Leach, M., Minot, E.O., Stewart, M., Stafford, K.J., 2016. Coding and quantification of a facial expression for pain in lambs. Behav. Process. 132, 49–56. https://doi.org/10.1016/j.beproc.2016.09.010.

Häger, C., Biernot, S., Buettner, M., Glage, S., Keubler, L.M., Held, N., Bleich, E.M., Otto, K., Müller, C.W., Decker, S., Talbot, S.R., Bleich, A., 2017. The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. PLoS One 12, e0175839. https://doi.org/10.1371/journal.pone.0175839.

Keating, S.C.J., Thomas, A.A., Flecknell, P.A., Leach, M.C., 2012. Evaluation of EMLA Cream for Preventing Pain during Tattooing of Rabbits: Changes in Physiological, Behavioural and Facial Expression Responses. PLoS ONE 7, e44437. https://doi.org/10.1371/journal.pone.0044437.

Knierim, U., C. Winckler, L. Mounier, and I. Veissier. 2021. Developing effective welfare measures for cattle. doi:10.19103/AS.2020.0084.05.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

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.

Lanci, A., Benedetti, B., Freccero, F., Castagnetti, C., Mariella, J., van Loon, J.P.A.M., Padalino, B., 2022. Development of a composite pain scale in foals: a pilot study. Animals 12, 439. https://doi.org/10.3390/ani12040439.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159–174. https://doi.org/10.2307/2529310.

Langford, D.J., Bailey, A.L., Chanda, M.L., Clarke, S.E., Drummond, T.E., Echols, S., Glick, S., Ingrao, J., Klassen-Ross, T., LaCroix-Fralish, M.L., Matsumiya, L., Sorge, R. E., Sotocinal, S.G., Tabaka, J.M., Wong, D., van den Maagdenberg, A.M.J.M., Ferrari, M.D., Craig, K.D., Mogil, J.S., 2010. Coding of facial expressions of pain in the laboratory mouse. Nat. Methods 7, 447–449. https://doi.org/10.1038/nmeth.1455.

Leach, M.C., Coulter, C.A., Richardson, C.A., Flecknell, P.A., 2011. Are we looking in the wrong place? Implications for behavioural-based pain assessment in rabbits (Oryctolagus cuniculi) and beyond. PLoS One 6, e13347. https://doi.org/10.1371/journal.pone.0013347.

Leung, V., Zhang, E., Pang, D.S., 2016. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats. Sci. Rep. 6, 31667. https://doi.org/10.1038/srep31667.

Luna, S.P.L., Trindade, P.H.E., Monteiro, B.P., Crosignani, N., della Rocca, G., Ruel, H.L. M., Yamashita, K., Kronen, P., Tseng, C.T., Teixeira, L., Steagall, P.V., 2022. Multilingual validation of the short form of the Unesp-Botucatu Feline Pain Scale (UFEPS-SF). PeerJ 10, e13134. https://doi.org/10.7717/peerj.13134.

McLennan, K.M., Rebelo, C.J.B., Corke, M.J., Holmes, M.A., Leach, M.C., Constantino-Casas, F., 2016. Development of a facial expression scale using footrot and mastitis as models of pain in sheep. Appl. Anim. Behav. Sci. 176, 19–26. https://doi.org/10.1016/j.applanim.2016.01.007.

Mialon, M.M., Deiss, V., Andanson, S., Anglard, F., Doreau, M., Veissier, I., 2012. An assessment of the impact of rumenocentesis on pain and stress in cattle and the effect of local anaesthesia. Vet. J. 194, 55–59. https://doi.org/10.1016/j.tvjl.2012.02.019.

Miller, A.L., Leach, M.C., 2015. The mouse grimace scale: a clinically useful tool. PLoS One. https://doi.org/10.1371/journal.pone.0136000.

Mogil, J.S., Pang, D.S.J., Silva Dutra, G.G., Chambers, C.T., 2020. The development and use of facial grimace scales for pain measurement in animals. Neurosci. Biobehav. Rev. 116, 480–493. https://doi.org/10.1016/j.neubiorev.2020.07.013.

Mokkink, L.B., Prinsen, C.A.C., Bouter, L.M., de Vet, H.C.W., Terwee, C.B., 2016. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Braz. J. Phys. Ther. 20, 105–113. https://doi.org/10.1590/bjpt-rbf.2014.0143.

Mokkink, L.B., C.A.C. Prinsen, D.L. Patrick, J. Alonso, L.M. Bouter, H.C.W. de Vet, and C. B. Terwee. 2018. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs).

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.

Orth, E.K., Navas González, F.J., Iglesias Pastrana, C., Berger, J.M., le Jeune, S.S., Davis, E.W., McLean, A.K., 2020. Development of a Donkey Grimace Scale to Recognize Pain in Donkeys (Equus asinus) Post Castration. Anim. (Basel) 10. https://doi.org/10.3390/ani10081411.

Parham, J.T., Tanner, A.E., Wahlberg, M.L., Grandin, T., Lewis, R.M., 2019. Subjective methods to quantify temperament in beef cattle are insensitive to the number and biases of observers. Appl. Anim. Behav. Sci.

Phythian, C.J., Toft, N., Cripps, P.J., Michalopoulou, E., Winter, A.C., Jones, P.H., Grove-White, D., Duncan, J.S., 2013. Inter-observer agreement, diagnostic sensitivity and specificity of animal-based indicators of young lamb welfare. Animal 7, 1182–1190. https://doi.org/10.1017/S1751731113000487.

Pinho, R.H., Leach, M.C., Minto, B.W., Rocha, F.D.L., Luna, S.P.L., 2020. Postoperative pain behaviours in rabbits following orthopaedic surgery and effect of observer presence. PLOS ONE 15, e0240605. https://doi.org/10.1371/journal.pone.0240605.

Pinho, R.H., Justo, A.A., Cima, D.S., Fonseca, M.W., Minto, B.W., Rocha, F.D.L., Leach, M.C., Luna, S.P.L., 2023. Effects of human observer presence on pain assessment using facial expressions in rabbits. J. Am. Assoc. Lab. Anim. Sci. 62, 81–86. https://doi.org/10.30802/AALAS-JAALAS-22-000056.

Reijgwart, M.L., Schoemaker, N.J., Pascuzzo, R., Leach, M.C., Stodel, M., de Nies, L., Hendriksen, C.F.M., van der Meer, M., Vinke, C.M., Van Zeeland, Y.R.A., 2017. The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. PLoS One 12, e0187986. https://doi.org/10.1371/journal.pone.0187986.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.

Sotocinal, S.G., Sorge, R.E., Zaloum, A., Tuttle, A.H., Martin, L.J., Wieskopf, J.S., Mapplebeck, J.C., Wei, P., Zhan, S., Zhang, S., McDougall, J.J., King, O.D., Mogil, J. S., 2011. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Mol. Pain. 7, 55. https://doi.org/10.1186/1744-8069-7-55.

Sprecher, D.J., Hostetler, D.E., Kaneene, J.B., 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. Theriogenology 47, 1179–1187.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293. https://doi.org/10.1126/science.3287615.

Tomacheuski, R.M., Monteiro, B.P., Evangelista, M.C., Luna, S.P.L., Steagall, P.V., 2023. Measurement properties of pain scoring instruments in farm animals: a systematic review using the COSMIN checklist. PLOS ONE 18, e0280830. https://doi.org/10.1371/journal.pone.0280830.

Torcivia, C., McDonnell, S., 2020. In-person caretaker visits disrupt ongoing discomfort behavior in hospitalized equine orthopedic surgical patients. Animals 10, 210. https://doi.org/10.3390/ani10020210.

Veissier, I., Mialon, M.M., Sloth, K.H., 2017. Early modification of the circadian organization of cows' activity in relation to disease or estrus. J. Dairy Sci. 100, 3969–3974. https://doi.org/10.3168/jds.2016-11853.

Vieira, A., Battini, M., Can, E., Mattiello, S., Stilwell, G., 2018. Inter-observer reliability of animal-based welfare indicators included in the Animal Welfare Indicators welfare assessment protocol for dairy goats. Animal 12, 1942–1949. https://doi.org/10.1017/S1751731117003597.

Williams, A.C., de, C., 2002. Facial expression of pain: an evolutionary account. Behav. Brain Sci. 25, 439–455. https://doi.org/10.1017/S0140525×02000080.

Yamada, P.H., Codognoto, V.M., de Ruediger, F.R., Trindade, P.H.E., da Silva, K.M., Rizzoto, G., Maestá, S.A., Ferreira, J.C.P., de Soutello, R.V.G., Oba, E., 2021. Pain assessment based on facial expression of bulls during castration. Appl. Anim. Behav. Sci. 236, 105258 https://doi.org/10.1016/j.applanim.2021.105258.

Zimov, J.L., Botheras, N.A., Weiss, W.P., Hogan, J.S., 2011. Associations among behavioral and acute physiologic responses to lipopolysaccharide-induced clinical mastitis in lactating dairy cows. Am. J. Vet. Res. 72, 620–627. https://doi.org/10.2460/ajvr.72.5.620.