The Veterinary Journal 304 (2024) 106079

Contents lists available at ScienceDirect

The Veterinary Journal

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

Refinement and revalidation of the Equine Ophthalmic Pain Scale: R-EOPS a new scale for ocular pain assessment in horses

S. Nannarone [a,b,c,*], F. Ortolani [a], N. Scilimati [a], R. Gialletti [a,b], L. Menchetti [d]

[a] Department of Veterinary Medicine, Veterinary Teaching Hospital, University of Perugia, Via San Costanzo 4, Perugia 06126, Italy
[b] Department of Veterinary Medicine, CRCS (Centro di Ricerca sul Cavallo Sportivo), University of Perugia, Via San Costanzo 4, Perugia 06126, Italy
[c] Department of Veterinary Medicine, CeRiDA (Centro di Ricerca sul Dolore Animale), University of Perugia, Via San Costanzo 4, Perugia 06126, Italy
[d] School of Bioscience and Veterinary Medicine, University of Camerino, Via Circonvallazione 93/95, Matelica 62024, Italy

ARTICLE INFO

Keywords:
Equine
Facial expressions
Ocular pain
Ophthalmic diseases
Pain measurement

ABSTRACT

This study addresses the refinement and revalidation of a composite pain scale that focuses on equine facial expressions and behavioural indicators as exhibitions of ophthalmic pain. This scale included only Behavioural and Facial and Ocular expression indicators and, compared to the first version of Equine Ophthalmic Pain Scale (EOPS), item descriptors and related ratings were changed. Thirteen horses with ocular diseases that required medical or surgical treatment were enroled (group P). In each animal, the refined EOPS (R-EOPS) was applied prior to any treatment (T0) and one week later (T7). The R-EOPS was applied twice, 7 days apart, to 16 healthy control horses (group C). Two 30-second videos were recorded each time to allow the retrospective analysis by eight observers. Inter-observer reliability of items was moderate or substantial (Krippendorff's alpha, Kα>0.40) while their intra-observer reliability was substantial or almost perfect for most items (Kα ≥0.61). Both inter- and intra-observer reliability of Total Score (TS) were however excellent (Intraclass Correlation Coefficients, ICC>0.75). The TS also showed good reproducibility (Kendall coefficient=0.786, ICC=0.684) and high consistency of its items (Cronbach's α=0.847). The comparison between groups as well as the sensitivity and specificity values supported the validity of the R-EOPS. In particular, for each extra point added to the TS, the risk of the horse having pain increased by more than two times (Odds Ratio=2.079, 95%CI=1.542–2.804; P<0.001). The Receiver Operating Characteristic analysis identified 6 as the threshold value of R-EOPS for discriminating horses with ocular pathology (sensitivity=83%, specificity=100%). This scale may be an effective tool for reliably assessing the pain level in horses with ophthalmic diseases and potentially guiding pain management although it still requires large-scale application and external validation.

Introduction

Recognition of pain in animals is clearly of significance for animal welfare, and the manifestation and experience of pain are influenced by several factors. In every species, pain is a complex multidimensional experience, conveying itself in behavioural, physiological, and emotional variables and their expressions (Prkachin, 2009).

Facial expressions are commonly studied and used to assess pain and other emotional states in humans, that have a prototypical 'pain face' (Prkachin, 2009). Facial expressions are evolutionarily preserved throughout mammalian species and, for this reason, they are easier for humans to identify and score also in animals (Descovich et al., 2017). Expressions result from the underlying facial musculature and related movements as immediate and spontaneous responses to pain (Ekman and Friesen, 1986). Thus, they could be useful as adjunct measures for evaluating animal welfare alongside existing indicators, such as behavioural rather than physiological parameters (Dalla Costa et al. 2014, AWIN, 2015).

Grimace scales have been developed to identify pain in animals and to potentially assess its severity in several animal species (Langford et al., 2010; Sotocinal et al., 2011; Gleerup et al., 2015a,b; Mogil et al., 2020; Orth et al., 2020; Van Dierendonck et al., 2020). They were found to be valid, with high inter-observer reliability scores ranging from 85% in horses (Dalla Costa et al., 2014) to 97% in piglets (Di Giminiani et al., 2016). The validity (i.e., does a scale measure what it claims to measure?) and reliability (i.e., the measurement error associated with a

* Corresponding author at: Department of Veterinary Medicine, Veterinary Teaching Hospital, University of Perugia, Via San Costanzo 4, Perugia 06126, Italy.
E-mail address: sara.nannarone@unipg.it (S. Nannarone).

https://doi.org/10.1016/j.tvjl.2024.106079
Accepted 8 February 2024
Available online 14 February 2024

scale) are in fact the key requirements for the successful application of pain scales (Mogil et al., 2020; Oliver et al. 2014). High reliability and validity imply that veterinarians and animal carers can assess pain in a systematic and consistent manner, ensuring that the animal receives proper care according to its painful condition (Richardson and Flecknell, 2005).

In particular, the Horse Grimace Scale (HGS) (Dalla Costa et al., 2014) has been applied to assess pain in horses after experimentally induced pain (Gleerup et al., 2015b), after castration (Dalla Costa et al., 2014), in dental-related pain (Marcantonio Coneglian et al., 2020), during musculoskeletal or orthopaedic pain (Dalla Costa et al., 2016; Dyson et al., 2017; van Loon and Van Dierendonck, 2019), colic syndrome (van Loon and Van Dierendonck, 2015; Van Dierendonck and van Loon, 2016) and in head-related pain (van Loon and Van Dierendonck, 2017). The HGS is also included in the Animal Welfare Indicators protocol for horses (AWIN, 2015). These studies showed that horse facial expressions and the HGS are effective methods for scoring soft tissue and orthopaedic pain in horses, but no study has specifically verified its applicability in horses affected by ocular pain.

Because the equine eye and especially the cornea are so prominent, they are very prone to traumatic injury and subsequent infection. Corneal ulceration or ulcerative keratitis is one of the most common and painful ocular problems for horses (Brooks and Plummer (2002)). Two important factors promote the overall health of the cornea: the tear film and the innervation (Knickelbein et al., 2018; Brooks and Plummer (2002)). Sensory innervation of the globe and adnexa is from the trigeminal nerve (cranial nerve V), and motor innervation is from the facial nerve (cranial nerve VII). The cornea is richly innervated and receives most of its sensory innervation from the terminal branches of the ciliary nerves, which arise from the ophthalmic division of the trigeminal nerve (Brooks and Plummer (2002)).

The outward evidence of equine ophthalmic diseases is obvious when blepharospasm, epiphora, eye rubbing, head tilt, asymmetry of shape or size compared to the unaffected eye, changes in the clarity of the cornea and noticeable abnormal discharges are evident (Brooks and Plummer (2002)). However, to the authors' knowledge, the degree of perceived pain has not been extensively evaluated yet, except in a preliminary context by our research group (Ortolani et al., 2021) and in a pilot study in experimental horses (Makra et al., 2021).

Our preliminary study (Ortolani et al., 2021) proposed a composite scale (Equine Ophthalmic Pain Scale, EOPS) including physiological, behavioural and specific facial expression indicators for assessing ocular pain. This study highlighted many strengths of the scale but also some weaknesses. In particular, our previous findings suggested that eliminating certain parameters from the EOPS, such as physiological indicators and some behavioural descriptors, would positively influence its accuracy and further shorten its application (Ortolani et al., 2021). Thus, in the present study, we hypothesise that the elimination of physiological parameters, the addition of further items related to facial expression, and the review of the scoring system could improve the accuracy and feasibility of the EOPS.

The objectives of this study were (1) to integrate the previous findings and to develop a new pain scoring system, named the Refined Equine Ophthalmic Pain Scale (R-EOPS), including the categories of behavioural and facial expressions for assessing pain related to ophthalmic diseases; (2) to test the new scale in horses affected by ocular or adnexa disease and in healthy horses; and (3) to validate the new scale by analysing its reliability, validity, sensitivity and specificity.

Materials and methods

The study was approved by the Bioethical Committee of XX (protocol number: 2019–16; Approval date, 1 July 2019). Informed owner consent was obtained for the inclusion of all horses in the study. Animals received a full physical and ophthalmic examination prior to the study period.

Sample size

The sample size calculation was based on Intraclass correlation coefficients (ICC) procedures using the confidence interval approach, as previously reported (Sutton et al., 2013; Menchetti et al., 2021). In particular, the following equation proposed by Machin et al. (2009) was used for the calculation of the sample size:

$$n = 1 + \frac{8 \times z^2_{1-\frac{\alpha}{2}} \times (1 - ICC_{plan})^2 \times [1 + (k-1) \times ICC_{plan}]^2}{k \times (k-1) \times W^2}$$

Where n= sample size, ICC_{plan}= planned ICC, k=raters, W= width of the confidence interval (CI). In the present study, a 90% CI of W=0.15 was chosen and an ICC_{plan}= 0.85 was planned. The minimum number of raters to be recruited was 4. The values for CI and W were chosen in consideration of the difficulty of finding animals that met the criteria required by the study. The required minimum total sample size was 24 horses.

Horses

Twenty-nine horses of different breeds, gender, coat colour and age were included. Sixteen horses found to be healthy and free of ocular and adnexa diseases after physical and ophthalmic evaluation were allocated to the control group (group C). They belonged to private riding schools and were housed in single horse boxes (3×3 m) on wood shavings, provided with water ad libitum and fed with hay four times a day. Thirteen horses admitted to the Veterinary Teaching Hospital of XX between January 2019 and September 2020 with ophthalmic and/or ocular adnexa diseases were included in the group of horses with ocular pathology (group P). Only animals older than one year, hospitalised for at least seven days, and filmed with good quality videos were included in the study. Mares with foal were excluded to limit disturbing effects due to mare-foal interaction during the R-EOPS assessment.

The Refined Equine Ophthalmic Pain Scale

For this study, a revised version of the previous EOPS, which in turn was adapted from pre-existing equine pain scales (Bussières et al., 2008; Dalla Costa et al., 2014; Gleerup et al. 2015b; Van Dierendonck and van Loon, 2016), was refined and its validity was checked. The new pain scale, R-EOPS, was developed through a combination of literature review, expert opinion and evaluation of the results of our previous trial (Ortolani et al., 2021). It is a multifactorial composite scale, which includes only two simple descriptive subscales, i.e., Behavioural and Ocular and Facial expression (OFex). Each subscale included specific items that were rated from 0 to 1 (indicating absence or presence, respectively) or from 0 to 2 (with 0 indicating normality and 2 corresponding to the most significant modification in the presence of pain). Scores within each subscale were summed, giving two partial scores (i.e., Behavioural and OFex partial scores); lastly, the total score (TS) was calculated as the sum of all the scores and ranged from 0 to 22, corresponding to the score identifying the absence of pain and maximal pain, respectively (Table 1).

The physiological parameters included in the previous EOPS were instead eliminated because they demonstrated poor consistency and extended application times (Ortolani et al., 2021).

Behavioural data subscale

Items included in this subscale, such as 'overall behaviour', 'position inside the box,' 'head position' and 'response to door opening' have already been reported in previous composite pain scales (Price et al., 2003; Pritchett et al., 2003; Bussières et al., 2008) and were also included in the first version of the EOPS (Ortolani et al., 2021). However, a different description and scoring were proposed for some of them (in particular for those that had previously shown lower reliability),

Table 1
Refined Equine Ophthalmic Pain Scale (R-EOPS). The scale includes 2 categories: 'Behavioural data sub-scale' and 'Ocular and facial expression sub-scale'.

Data scale		Score
Behavioural data sub-scale		
Overall behaviour*	Quietly standing/Looking for food	0
	Depress/nervous	1
	Lying down	2
Position inside the box*	In front of the box, observing the environment/looking at the door or looking for food	0
	In the middle of the box, looking at the side	1
	Giving the back to the door	2
Head position (occipital region)	Above the withers	0
	At the level of the withers	1
	Below the withers	2
Response to door opening*	Coming closer/ approaching/ staring at the door	0
	Turning the face in the other side/ walking away	1
Partial score 1		…/7
Ocular and Facial expression sub-scale (Ofex)		
Ears movements	Both ears facing forward	0
	Both ears are moving in different directions or placed in an asymmetrical position	1
	Both ears facing back	2
Tension above the eye area*	Not present	0
	Moderately present	1
	Obviously present	2
Blepharospasm	Not evident	0
	Slightly evident	1
	Evident	2
Prominent strained chewing muscle*	Not present	0
	Moderately present	1
	Obviously present	2
Mouth strained and pronounced chin*	Not present	0
	Moderately present	1
	Obviously present	2
Straining nostrils and flattening of the profile*	Not present	0
	Moderately present	1
	Obviously present	2
Lacrimation	Not present	0
	Present	1
Response to eyelids opening	No response	0
	Resistance	1
	Avoiding physical contact	2
Partial score 2		…/15
Total Score (TS)		…/22

[*] Items that have descriptors modified with respect to EOPS.

with the aim of improving their interpretability and reducing ambiguity (Streiner and Norman, 2008). In particular, the authors tried to include more straightforward descriptors, and eliminated 3 as the highest score, according to the previous results which highlighted the redundancy of the 4-point scales (Ortolani et al., 2021). As a result, the maximum value of this partial score was 7/22. It was lower than that in EOPS (11/31), but its weight on the TS remained the same (i.e., about one third of the TS).

Ocular and Facial expression subscale (OFex)

Unlike the EOPS, HGS descriptors such as 'tension above the eye area', 'prominent strained chewing muscles', 'mouth strained and pronounced chin' and 'strained nostrils and flattening of the profile' have been introduced in this new composite scale according to Dalla Costa et al. (2014). This led to the decision to modify the name from OcEx of the EOPS, into OFex, given the presence of both ocular and facial items. The items 'lacrimation' and 'response to eyelids opening,' which showed a good correlation with the TS (ρ >0.3) and improved the consistency of the previous EOPS (Ortolani et al., 2021), were maintained in the new R-EOPS. The evaluation of the 'response to eyelids opening' was performed only by the independent observer on-site, who interacted with each animal and could not film this assessment for the subsequent evaluation by the eight blinded observers. Indeed, as described for the EOPS development, the independent observer, who filmed all animals on its own, could not open the eyelids and video record the reaction at the same time, given the need to hold the halter with the other hand (Ortolani et al., 2021). The OFex subscale partial score increased to a maximum of 15/22, compared to 7/31 in the EOPS, leading to an increase of weight on the TS of about two thirds.

Video recording and pain scoring

Animals were filmed two times to assess and score pain: at the time of admission (T0), before any type of sedation or treatment, for horses in group P, or at the baseline for group C; and seven days later (T7) in both groups. For each time point (i.e., T0 and T7), two videos lasting 30 s each were recorded by an independent observer using a smartphone (iPhone®). The first video filmed the horse while it was undisturbed inside its box up to the time the observer opened the box door to assess the Behavioural scale; the second video focused on the head profile, including ears and muzzle, to assess and rate the OFex scale. For this assessment, the horse was kept without the halter. No ocular catheter or bandages were present at the time of video recording to reduce possible bias for the observers. Lighting conditions were not controlled but were based on the ambient light present at the time of video recording.

A total of eight observers (5 veterinarians and 3 recently graduated veterinarians) were recruited (the minimum number that had been envisaged was 4 raters). Among them, the most experienced in equine medicine and behavioural assessment (main observer) trained the other observers in pain scoring evaluation, including practical application of the EOPS to animals. As soon as the observers became confident, they all independently scored the Behavioural data scale and OFex scale (except for 'response to eyelids opening'). They were all blinded to both the group and time when the videos were recorded, and they independently scored 116 videos from 29 horses (group C, n=16; group P, n=13) recorded at T0 and T7. The videos were numbered using a random sequence generator.

To obtain the TS for each animal, the scores assigned by the main observer were added to the score for 'response to eyelids opening' recorded by the independent observer, the only one who was able to assign the score for this item. Furthermore, the main observer re-evaluated the videos two months apart, with a different order and identification number, to calculate intra-observer reliability. The independent observer, who filmed the videos and assigned *in vivo* scores, was not involved in video assessments, as he was not blinded to the treatment group.

Statistical analysis

The validation of the proposed scoring system included the following analyses: descriptive statistics, inter- and intra-observer reliability (reliability); test-retest reliability (reproducibility); internal consistency and item-total correlation; construct and criterion validity (Field et al., 2009; Meagher, 2009; Boateng et al. 2018; Menchetti et al., 2019; Silva et al., 2020). Table 2 summarises the statistical techniques and criteria used. For more details, see Ortolani et al., 2021. Tests for independent samples were used as no matching criterion between P and C horses could be found.

The data were recorded on an Excel spreadsheet (Excel 2007 Microsoft Corporation, Redmond, WA, USA) and then transferred to SPSS software version 25 (SPSS Inc., Chicago, IL) for statistical analysis. P values <0.05 were considered statistically significant.

Table 2
Statistical methods used for validation of the refined Equine Ophthalmic Pain Scale (R-EOPS) applied in healthy (C) and horses with ocular pathology (P) assessed at time 0 (prior to any treatment for group P; T0) and after one week (T7).

Type of analysis	Statistical test	Criteria
Distribution of scores	Descriptive statistics	Presentation of medians, interquartile range (IQR), number and percentage
Inter and intra-observer reliability	Krippendorff's alpha (Kα; using 1000 bootstrap samples to estimate the 95% confidence interval)[b] for individual items and Intraclass correlation coefficients (ICC; using the two-way ANOVA approach for single measurement and absolute agreement type[c]) for TS[d].	The Krippendorff's alpha (Kα) coefficient was interpreted as none to slight ($0.01\leq$ Kα < 0.20), fair ($0.21 \leq$ Kα < 0.40), moderate ($0.41\leq$ Kα < 0.60), substantial ($0.61\leq$ Kα $<$ 0.80), and almost perfect (Kα ≥ 0.81) agreement[f]. ICC values were interpreted as poor (ICC < 0.40), fair ($0.40 \leq$ ICC < 0.60), good ($0.60 \leq$ ICC<0.75), and excellent (ICC ≥ 0.75)[g,h].
Test-retest reliability	Kendall correlation coefficient tau (τ) and ICC.	Consistent and agreement between results conducted at two different times (T0 and T7) was calculated for scores obtained in group C[i,j]. The τ could range from 0 (no concordance) to 1 (perfect concordance). Associations were considered as weak if $\tau < 0.30$, moderate if $0.30 \leq \tau \leq 0.50$, and strong if $\tau > 0.50$[k]. For the ICC the same rules described above were used[g,h].
	Wilcoxon signed-rank tests	Assess whether there are differences in TS between T0 and T7 in the C group.
Internal consistency and corrected[a] item-total correlation	Cronbach's alpha (α)[e]	Values of $\alpha > 0.76$ were considered acceptable[l].
	Spearman correlation coefficient (ρ)	On a reliable scale, each item should have a $\rho > 0.30$ with the TS[l].
Construct validity	Mann–Whitney *U*	Assess whether there are differences in TS between healthy and pathologic horses (responsiveness between groups). The hypothesis was that the scale should increase if there is a painful stimulus[m] such as ocular pathology.
	Wilcoxon signed rank tests	Assess whether there are differences in TS between T0 and T7 in the P group (responsiveness over time). The hypothesis was that the scale should decrease after medical/surgical treatment and over time[m].
Criterion validity	Generalised linear model using a binomial distribution and logit as function link	Generalised Estimating Equations procedures were used to estimate the Predictive validity[n]. Horse ID was included as subject variable, Time as within-subject variable (assuming exchangeable working correlation matrix), and the TS was treated as a continuous predictor. Results were expressed as odds ratio (OR) with 95% CI and the *P* value. The hypothesis was that the odds that an ocular pathology will occur increase for each 1-unit increase in TS.
	Receiver operating characteristic (ROC) analysis[e]	The presence of ocular pathology was set as the positive actual state and larger values of the test result variable indicated stronger evidence for a positive actual state. Based on the statistics of the AUC, the R-EOPS may be considered as uninformative (AUC = 0.50), poorly accurate ($0.50 \leq$ AUC $\leq$ 0.70), moderately accurate ($0.70 \leq$ AUC ≤ 0.90), very accurate ($0.90 \leq$ AUC<1) or perfect (AUC = 1). The optimal cut-off was determined as the point of the curve closest to (0,1), i.e., Youden's index[o].

ANOVA = Analysis of variance; TS = total score; C=healthy horses; P= horses with ocular pathology; AUC = area under the curve.

[a] The value of the target item is subtracted from the total
[b] Krippendorff, 1970
[c] McGraw and Wong, 1996
[d] 'response to eyelids opening' was not included in this TS as only recorded by the independent observer
[e] Calculated with the data collected at T0
[f] McHugh, 2012; Tallon et al., 2021
[g] Fleiss, 1986
[h] Hallgren, 2012
[i] Meagher, 2009
[j] Menchetti et al., 2019
[k] Cohen, 1988
[l] Field et al., 2009
[m] Silva et al., 2020
[n] Boateng et al., 2018
[o] Greiner et al., 2000

Results

Population

Twenty-nine horses, both healthy (group C, *n*=16) and horses with ocular pathology (group P, *n*=13), were included in the study. Data of one horse of group P were not completed as two videos were excluded due to poor-quality images. There were eighteen mares, ten castrated males and one intact male of different breeds: Arabian (*n*=3), Maremmano (*n*=3), Standardbred (*n*=2), Thoroughbred (*n*=2), KWPN (*n*=1), Hannover (*n*=13), warmblood (*n*=4), Polo pony (*n*=1). The mean age ± standard deviation was 10.6 ± 6.2 years (range 1–24 years). Animals in group P were affected with different ocular diseases: corneal ulceration (*n*=5), iris prolapse + corneal ulceration (*n*=1), nictitating membrane habronemiasis + corneal ulceration (*n*=1), anterior uveitis + corneal ulceration (*n*=1), pseudomonas ulcerative keratitis (*n*=1), mycotic ulcerative keratitis (*n*=1), lens luxation (*n*=1), recurrent uveitis (*n*=1) and corneal stromal abscess (*n*=1). They received medical treatment appropriate for each case, which could include topical antibiotics, antimycotics and cycloplegics, administered several times a day, alone or associated with surgical treatment. Moreover, flunixin meglumine 1 mg/kg was administered intravenously once or twice as an anti-inflammatory and analgesic drug, regardless of the R-EOPS score.

Distribution of scores

The distribution of scores according to group and observation time is given in Table 3. Fig. 1 shows an example of horses with ophthalmic disease and is representative of some drawbacks encountered during R-EOPS application, specifically due to different breeds or coat colour and length.

Inter and intra-observer reliability

Inter-observer reliability of each item evaluated by Krippendorff's alpha coefficient was moderate (at one or both time points) for 'overall behaviour', 'tension above the eye area', 'prominent strained chewing muscles', 'mouth strained and pronounced chin', and 'strained nostrils and flattening of the profile' ($0.41 \leq K\alpha < 0.60$) while it was substantial ($0.61 \leq K\alpha < 0.80$) for the other items (Table 4). Intra-observer reliability was substantial or almost perfect for most items ($K\alpha \geq 0.61$; Table 5). A Kα value less than 0.6 (indicating moderate Intra-observer reliability) was only found for 'tension above the eye area' at T0.

The reliability of the TS was evaluated by ICC. Both inter-observer (T0: ICC=0.884, 95%CI=0.817–0.936, $p<0.001$; T1: ICC=0.876, 95% CI=0.801–0.933, $p<0.001$) and intra-observer (T0: ICC=0.928, 95% CI=0.850–0.966, $p<0.001$; T1: ICC=0.915, 95%CI=0.822–0.960, $p<0.001$) reliability of the TS was excellent.

Test-retest reliability and differences between T0 and T7 in group C

The TS of healthy animals (group C) showed strong concordance (τ=0.786, $P<0.001$), good agreement (ICC=0.684), and no difference between the first and the second assessment (Mdn=1.5 and 1.0, IQR=0.0–5.0 and 0.0–4.00 at T0 and T7, respectively; P=0.317).

Internal consistency and item-total correlation

Cronbach's α (0.847) indicated a good internal consistency of the R-EOPS. An improvement in the α value would be obtained by removing 'head position' (Cronbach's α=0.861), 'response to door opening' (Cronbach's α=0.857) and 'prominent strained chewing muscles' (Cronbach's α=0.860). These indicators were also poorly correlated with the TS ($\rho<0.30$). Only a very little improvement would be obtained by eliminating 'overall behaviour' (Cronbach's α=0.848; Table 6).

Construct and criterion validity, and clinical application

The difference between TSs for healthy horses and horses with an ophthalmic disease was significant both at T0 ($P<0.001$) and at T7 (P=0.005; Fig. 2). For each extra point added to the TS, the risk of the horse being affected by ocular pathology increased by more than two times (OR=2.079, 95%CI=1.542–2.804; $P<0.001$). The higher TS of

Table 3
Distribution of scores for each item in healthy (C) and horses with ocular pathology (P) assessed at time 0 (prior to any treatment for P group; T0) and after 1 week (T7). Values are number (*n*) and percentage (%) for each score and medians and IQR for partial scores.

Category	Item	Score	Time			
			T0		T7	
			Group		Group	
			C *n* (%)	P *n* (%)	C *n* (%)	P *n* (%)
Behavioural data scale	Overall behaviour	0	14 (87.5%)	9 (75.0%)	15 (93.8%)	10 (83.3%)
		1	2 (12.5%)	3 (25.0%)	1 (6.3%)	2 (16.7%)
	Position inside the box	0	16 (100.0%)	5 (41.7%)	16 (100.0%)	7 (58.3%)
		1	0 (0.0%)	4 (33.3%)	0 (0.0%)	4 (33.3%)
		2	0 (0.0%)	3 (25.0%)	0 (0.0%)	1 (8.3%)
	Head position	0	13 (81.3%)	9 (75.0%)	15 (93.8%)	8 (66.7%)
		1	3 (18.8%)	1 (8.3%)	1 (6.3%)	4 (33.3%)
		2	0 (0.0%)	2 (16.7%)	0 (0.0%)	0 (0.0%)
	Response to door opening	0	15 (93.8%)	10 (83.3%)	16 (100.0%)	11 (91.7%)
		1	1 (6.3%)	2 (16.7%)	0 (0.0%)	1 (8.3%)
Behavioural scale partial score (median and IQR)			0 (0–1)	1 (0–4)	0 (0–0)	1 (0–2)
Ocular and Facial Expression scale (OFex)	Ears movements	0	4 (25.0%)	1 (7.7%)	5 (31.3%)	4 (30.8%)
		1	11 (68.8%)	9 (69.2%)	11 (68.8%)	7 (53.8%)
		2	1 (6.3%)	3 (23.1%)	0 (0.0%)	2 (15.4%)
	Tension above the eye area	0	14 (87.5%)	4 (30.8%)	13 (81.3%)	5 (38.5%)
		1	2 (12.5%)	9 (69.2%)	3 (18.8%)	5 (38.5%)
		2	0 (0.0%)	2 (15.4%)	0 (0.0%)	3 (23.1%)
	Blepharospasm	0	16 (100.0%)	4 (30.8%)	16 (100.0%)	5 (38.5%)
		1	0 (0.0%)	4 (30.8%)	0 (0.0%)	3 (23.1%)
		2	0 (0.0%)	5 (38.5%)	0 (0.0%)	5 (38.5%)
	Prominent strained chewing muscle	0	12 (75.0%)	6 (46.2%)	11 (68.8%)	8 (61.5%)
		1	4 (25.0%)	7 (53.8%)	5 (31.3%)	3 (23.1%)
		2	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (15.4%)
	Mouth strained and pronounced chin	0	14 (87.5%)	2 (15.4%)	16 (100.0%)	5 (38.5%)
		1	2 (12.5%)	10 (76.9%)	0 (0.0%)	7 (53.8%)
		2	0 (0.0%)	1 (7.7%)	0 (0.0%)	1 (7.7%)
	Straining nostrils and flattening of the profile	0	14 (87.5%)	5 (38.5%)	14 (87.5%)	7 (53.8%)
		1	2 (12.5%)	7 (53.8%)	2 (12.5%)	5 (38.5%)
		2	0 (0.0%)	1 (7.7%)	0 (0.0%)	1 (7.7%)
	Lacrimation	0	16 (100.0%)	5 (38.5%)	16 (100.0%)	9 (69.2%)
		1	0 (0.0%)	8 (61.5%)	0 (0.0%)	4 (30.8%)
	Response to eyelids opening	0	15 (93.8%)	2 (15.4%)	12 (75.0%)	4 (30.8%)
		1	1 (6.3%)	9 (69.2%)	4 (25.0%)	7 (53.8%)
		2	0 (0.0%)	2 (15.4%)	0 (0.0%)	2 (15.4%)
OFex partial score (median and IQR)			1 (1–2)	8 (4–9)	1 (1–2)	5 (2–9)

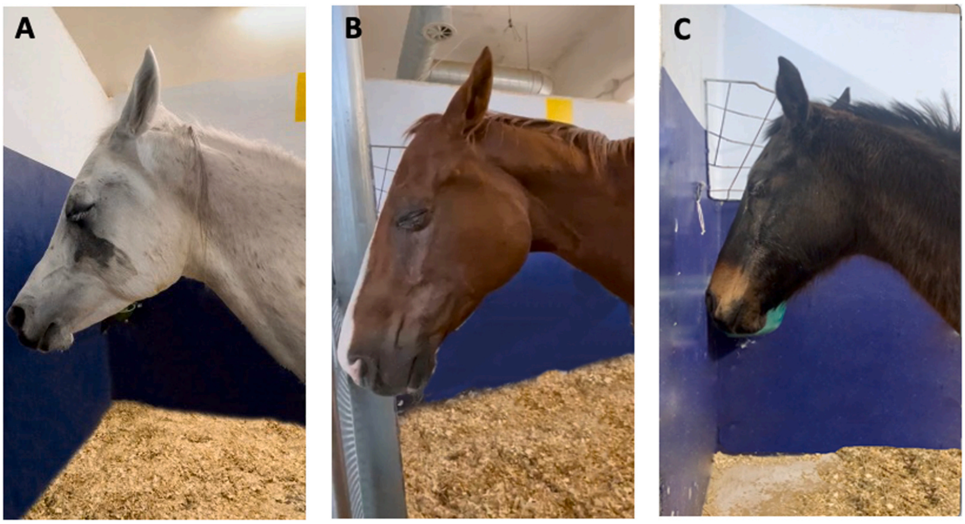

Fig. 1. Example of three horses with ocular pathology representative of possible drawbacks when applying the R-EOPS due to different breed or coat colour and length. An example of scoring is provided beside each indicator in brackets. (A) Arabian horse, note the edge-shaped head, pointed ears and smaller muzzle respect to B and C. The R-EOPS results in evident blepharospasm (2), lacrimation (1), both ears facing backward (2), strained nostrils and flattening of the profile obviously present (2), prominent strained chewing muscle moderately present (1), and mouth strained and pronounced chin moderately present (1). (B) Warmblood horse with short hair and bright colour coat showing evident blepharospasm (2), both ears facing backward (2), strained nostrils and flattening of the profile moderately present (1), prominent strained chewing muscle obviously present (2), and tension above the eye area moderately present (1). (C) Some items included in the Ocular and Facial Expression scale are more difficult to assess in this horse respect to A and B, because of the hairy and dark coat. However, note the evident blepharospasm (2), lacrimation (1), both ears placed in an asymmetrical position (1), and prominent strained chewing muscle moderately present (1).

Table 4
Inter-observer reliability. Krippendorff's alpha (Kα) of items included in the Behavioural and OFex categories assessed by eight observers. Each Kα is followed by its 95% confidence interval (CI).

Item	Time	Kα	95% CI	
			Lower bound	Upper bound
Overall behaviour	T0	0.586	0.500	0.658
	T7	0.589	0.476	0.681
Position inside the box	T0	0.768	0.725	0.816
	T7	0.819	0.770	0.863
Head position	T0	0.606	0.543	0.667
	T7	0.609	0.536	0.675
Response to door opening	T0	0.745	0.662	0.821
	T7	0.696	0.587	0.804
Ears movements	T0	0.655	0.613	0.693
	T7	0.646	0.598	0.691
Tension above the eye area	T0	0.574	0.515	0.635
	T7	0.613	0.558	0.663
Blepharospasm	T0	0.928	0.909	0.946
	T7	0.866	0.835	0.895
Prominent strained chewing muscles	T0	0.449	0.379	0.510
	T7	0.565	0.498	0.625
Mouth strained and pronounced chin	T0	0.502	0.442	0.554
	T7	0.479	0.411	0.552
Strained nostrils and flattening of the profile	T0	0.585	0.523	0.644
	T7	0.479	0.405	0.552
Lacrimation	T0	0.879	0.841	0.911
	T7	0.652	0.564	0.733

The Kα was interpreted as none to slight (0.01≤ Kα < 0.20), fair (0.21 ≤ Kα < 0.40), moderate (0.41≤ Kα < 0.60), substantial (0.61≤ Kα < 0.80), and almost perfect (Kα ≥ 0.81) agreement.

Table 5
Intra-observer reliability. Krippendorff's alpha (Kα) of items included in the Behavioural and OFex categories assessed by eight observers. Each Kα is followed by its 95% confidence interval (CI).

Item	Time	Kα	95% CI	
			Lower bound	Upper bound
Overall behaviour	T0	0.610	0.090	1.000
	T7	0.840	0.519	1.000
Position inside the box	T0	0.994	0.983	1.000
	T7	0.897	0.696	1.000
Head position	T0	0.991	0.973	1.000
	T7	0.766	0.418	1.000
Response to door opening	T0	1.000	1.000	1.000
	T7	1.000	1.000	1.000
Ears movements	T0	0.607	0.401	0.793
	T7	0.749	0.546	0.925
Tension above the eye area	T0	0.571	0.516	0.620
	T7	0.623	0.578	0.665
Blepharospasm	T0	0.893	0.757	0.992
	T7	0.936	0.822	1.000
Prominent strained chewing muscles	T0	0.633	0.340	0.853
	T7	0.860	0.662	1.000
Mouth strained and pronounced chin	T0	0.776	0.579	0.930
	T7	0.841	0.613	1.000
Strained nostrils and flattening of the profile	T0	0.730	0.420	0.972
	T7	0.740	0.406	0.994
Lacrimation	T0	0.918	0.755	1.000
	T7	1.000	1.000	1.000

The Kα was interpreted as none to slight (0.01≤ Kα < 0.20), fair (0.21 ≤ Kα < 0.40), moderate (0.41≤ Kα < 0.60), substantial (0.61≤ Kα < 0.80), and almost perfect (Kα ≥ 0.81) agreement.

animals in painful conditions supports the responsiveness between groups of the R-EOPS. The TS of horses with an ophthalmic disease did not change significantly over time (P=0.208), thus not supporting overtime responsiveness, but increased its variability at T7 compared to T0 (median=8 and 6, IQR=6–12 and 2–10 at T0 and T7, respectively).

The AUC analysis showed that the R-EOPS was a very accurate method for discriminating between healthy horses and those with ophthalmic disease (AUC=0.951, 95%CI= 0.873–1.000; P<0.001; Fig. 3). Its optimal threshold value (optimal cutoff) was 6 (i.e., positive if greater than or equal to 6), which made it possible to obtain 83% sensitivity and 100% specificity (Table 7).

Table 6
Parameters indicating internal consistency: correlation (ρ) between each item and the Total Score (TS; on a reliable scale, each item should have a ρ >0.3 with the TS), and Cronbach's α if each item was deleted from the composite scale. Cronbach α of the full scale (*n* items = 12) was 0.847.

Item	Corrected[a] item-TS correlation (ρ)	Cronbach's alpha if item deleted
Overall behaviour	0.306	0.848[b]
Position inside the box	0.628	0.827
Head position	0.182	0.861[b]
Response to door opening	0.084	0.857[b]
Ears movements	0.466	0.839
Tension above the eye area	0.702	0.823
Blepharospasm	0.734	0.816
Prominent strained chewing muscles	0.128	0.860[b]
Mouth strained and pronounced chin	0.756	0.817
Strained nostrils and flattening of the profile	0.646	0.828
Lacrimation	0.844	0.816
Response to eyelids opening	0.704	0.820

[a] The value of each item was subtracted from the TS.
[b] α values that would improve by eliminating the corresponding items

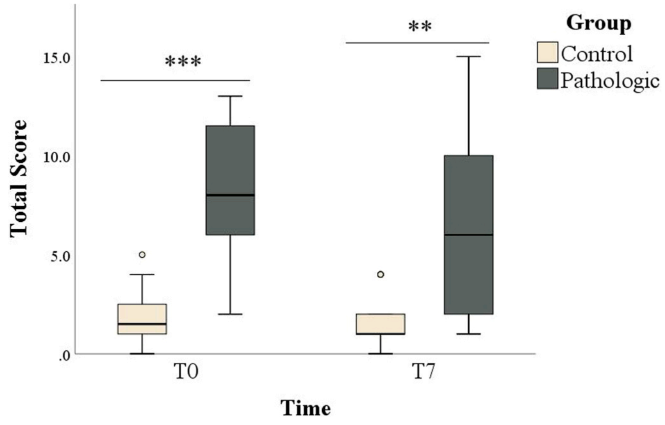

Fig. 2. Box plots of R-EOPS TS scores in healthy horses (group C, light grey) and horses with ophthalmic disease (group P, dark grey) at baseline (T0) and seven days later (T7). Asterisks indicate the significance of the differences between groups (C *versus* P) within each time (***$P<0.001$; **$P<0.01$).

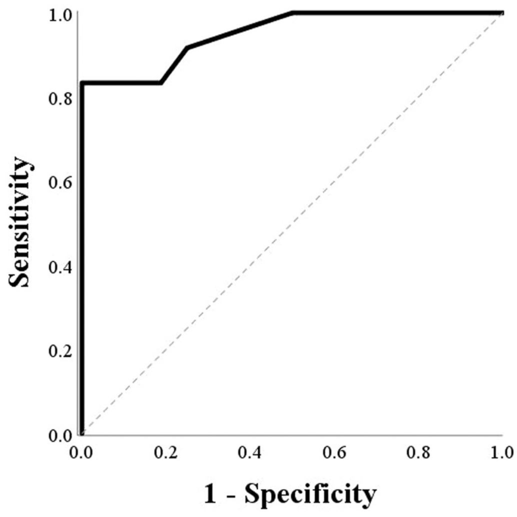

Fig. 3. Receiver Operating Characteristic (ROC) curve of TS in discriminating healthy from animals with ophthalmic disease (AUC=0.951, 95%CI= 0.873–1.000; $P<0.001$). The dotted line indicated the reference line (AUC=0.5). A score of 6 resulted as optimal value for discriminating healthy horses from horses with an ophthalmic disease based on the balance of highest sensitivity (83%) and specificity (100%). 95% CI, 95% confidence intervals.

Table 7
Sensitivity and specificity (with respective 95% CI) associated with some potential Total Scores.

Total Score*	Sensitivity (100%)	95% CI	Specificity (100%)	95% CI
1	100	46–100	13	13–13
2	100	46–100	50	50–50
3	92	42–100	75	40–100
4	83	39–100	81	43–100
5	83	39–100	94	48–100
6	83	39–100	100	51–100
7	58	36–81	100	51–100
9	42	42–42	100	51–100
10	33	33–33	100	51–100
12	25	25–25	100	51–100
13	8	8–8	100	51–100
14	0	0–0	100	51–100

A score of 6 resulted as the optimal value for discriminating healthy horses from horses with an ophthalmic disease based on the balance of the highest sensitivity (83%) and specificity (100%). 95% CI, 95% confidence intervals.
* Positive if Greater Than or Equal To.

Discussion

This study assessed the validity of a refined version of the Equine Ophthalmic Pain Scale (R-EOPS) for assessing and scoring the presence of pain in horses with ocular or adnexa diseases. The present study, in particular, supported (1) the inter- and intra-rater reliability as well as test-retest reliability, (2) the internal consistency, (3) the predictive validity, and (4) the construct validity of R-EOPS although its responsiveness over time and to analgesic treatment could not be demonstrated. Moreover, a cutoff score to discriminate healthy horses from horses with an ophthalmic disease was proposed. The R-EOPS is a composite scale that includes a facial expression scale, integrated with items strictly related with ophthalmic symptoms, and a behavioural scale. It seems a useful and reliable tool but also leaves room for improvement, paving the way for further investigations on a larger sample size and in real-time, and for external validation.

The inter-observer agreement of the TS of R-EOPS, as well as its intra-observer agreement, was excellent. This finding supports its overall reliability. However, the inter-observer agreement of the individual items analysed with the approach suggested by Krippendorff (1970) revealed some weaknesses. In particular, 'overall behaviour', 'tension above the eye area', 'prominent strained chewing muscles', 'mouth strained and pronounced chin', and 'strained nostrils and flattening of the profile' only showed a moderate agreement (Kα <0.60). A direct comparison with the previous EOPS is unlikely given the different statistical tools used in this refinement process. However, these low inter-observer agreements could result from inadequate training of the observers, especially for those items that could be difficult to properly describe and score, such as those of the OFex. A similar drawback was described also for the HGS, where the items *tension above eye*, *strained mouth and pronounced chin* and *prominent strained chewing muscles* resulted as 'not able to be scored' in 15–21% of images (Dalla Costa et al., 2014). Pilot observations demonstrated that pain scoring from video recordings resulted in less detailed observation of facial expressions and lower inter-observer reliability (van Loon and Van Dierendonck, 2017). During video assessment, the observer could be influenced by excessive movements, such as those of the ears due to environmental noise, or by chewing, and likewise, the masseter tension has been considered as an indicator of acute stress in horses (Rankins et al., 2022). It is therefore likely that some items of R-EOPS are yet too subjective in their evaluation especially if assessed through video rather than still images.

Some animal-related factors could also influence the characteristics of the facial expressions resulting in confounders. For example, the muscular tone could change during the ageing process, therefore apparently different chewing muscles or greater tension above the eye could be simply related to a physiological change in elderly horses. However, we did not investigate a possible correlation between age and items with moderate inter-observer reliability. As regards the 'prominent strained chewing muscles', it should be also noted that in some horses it was impossible to record videos without the halter due to their uncooperative behaviour. Since the halter's sideband rests upon the chewing muscles, an important bias is likely to have occurred in these horses. Thus, the influence of the animal's temperament and the presence of the halter should be taken into consideration.

Finally, it is worth mentioning the technical issues related to the agreement index used to estimate the reliability of individual items. In the present study, Krippendorff's Alpha was used, an appropriate index of reliability for scales of measurement that however may suffer from some paradoxical behaviours (i.e., Cohen's k paradox) leading to an underestimation of the real agreement (Giammarino et al., 2021; Gwet, 2002, 2015).

The R-EOPS of healthy horses showed strong concordance between the first and the second assessment, demonstrating the reproducibility of the scale. Moreover, the scale developed in this study showed an improved internal consistency (R-EOPS Cronbach's α =0.85) compared to its previous version (EOPS Cronbach's $\alpha = 0.76$). The exclusion of physiological parameters, the refinement of the description of behavioural items, the simplification of the scoring system, and the introduction of new items in the Ofex subscale may have contributed to the increase in consistency. Nevertheless, a small improvement in the α value would be further obtained by removing 'head position', 'response to door opening' and 'prominent strained chewing muscles' parameters. These findings could arise from the difficulty of an objective assessment, as already evidenced by the inter-observer reliability results, and/or from a low variability of the distribution of their scores. The item 'response to door opening', for example, received a score of 0 in most of the horses, regardless of the group and time of assessment. This item, therefore, had little discriminating value and, for this reason, it correlated little with the TS (which instead, as discussed later, showed a good ability to discriminate horses with ophthalmic pain). This finding could suggest that although horses were affected by ophthalmic disease, pain may have not been severe enough to induce a relevant central depression that would have prevented a reaction to the environment, and eventual socialisation, in response to an attempted interaction after opening the door.

The TS results from the sum of all indicators included in both the Behavioural and Ofex subscales. The partial scores of the two subscales could be also examined separately to investigate which, among the behavioural or physical responses, best reflects the severity of the clinical pain. The validity of R-EOPS was nevertheless checked using the TS values. Like in the previous version (Ortolani et al., 2021), there were significant differences between the TS of horses in groups C and P at both observational times, and this supports the R-EOPS responsiveness as well as its discrimination capacity. However, in this horse population, the TS of horses in group P did not change significantly over time but only showed an increase in its variability at T7. This could suggest that only some horses were clinically improved after one week, thus reducing their TS, while others did not substantially change their apparent health status and had maintained a high TS. This is not surprising as it has been demonstrated that some corneal ulcerations (the most frequent disease in our population, with 11 out of 13 enroled horses affected) might require more than 15 days for complete healing, likely including a persisting painful condition seven days after medical and/or surgical treatment (Lassaline-Utter et al., 2014; Prucha et al., 2020). Since we are not sufficiently confident that pain had reduced in most of our horses at the second R-EOPS application, the responsiveness over time cannot be taken into consideration for the validation procedures. In fact, over time responsiveness assumes that a real change in the construct occurs (Mokkink et al., 2021). Thus, to adequately evaluate this aspect of the construct validity, the scale should be applied by stratifying animals according to their pathology and repeating the scoring after a longer period to consider the different times necessary for disease resolution and for the analgesic effectiveness.

We should mention again that the scale was not applied to guide the timing of analgesic treatment in the enroled horses affected by the ophthalmic disease, which relied upon the ophthalmologist's decision based on standard clinical practice. However, a cutoff of 6 was identified as the TS threshold for discriminating healthy horses from horses with an ophthalmic disease, which are likely to perceive pain. This cutoff proved to be sensitive (83%) and very specific (100%); therefore, the R-EOPS could find future clinical applications and suggest analgesic requirements.

This study has some limitations, such as the inclusion of Arabian horses, which have important differences in head morphology compared to other breeds. This could have impaired the evaluation of some Ofex items, particularly 'tension above the eye area,' 'prominent strained chewing muscle' and 'straining nostrils and flattening of the profile'. Indeed, the Arabian horse has an extremely refined, wedge-shaped head, a broad forehead, small ears, large eyes, large nostrils, and a small muzzle; it mostly has a distinctive concave, or 'dished' profile (Taha et al., 2017). Similarly, it has been shown that the colour of the horse's coat can interfere with scoring. Dark horses could be more difficult to assess than those with brighter coats, especially if filmed against a dark background (Dalla Costa et al., 2014). The reliability of some items could thus benefit from a better quality of their description, even providing hand-labelled drawings and/or pictures that highlight each area of interest and are specific for the different breeds and for each score. This could focus the observer's attention on the anatomical area to be evaluated and reduce subjective interpretation.

The horse's temperament should also be considered during any practical applications of the R-EOPS, as it may influence its behaviour and coping strategies, and thus its clinical manifestations of pain and distress (Menchetti et al., 2021; Riva et al., 2022). Additionally, the different environmental settings could create a bias. Despite attempts to provide similar conditions with regard to box size and feeding management, horses in group P were not in their usual environment, a stressor that could influences behaviour and facial expressions. Moreover, some boxes in horses of group C (n=4) had a window that could have influenced the interactive behaviour when the operator approached the horse and created a bias for the observer. However, these windows did not appear in the video-recording thus leaving blinded the observer when evaluating the clips.

As mentioned above, it is likely that observers of the current study did not receive adequate training. A recent study evidenced that observers of the HGS showed low inter-rater agreement, with a 30-minute training session being insufficient for inexperienced raters to obtain satisfactory inter-rater agreement (Dai et al., 2020). Nevertheless, the development of a standardised training protocol could improve the reliability of the pain rater. Moreover, to establish that the scale is valid under field conditions, clinical validation with less experienced observers should be accomplished.

Regarding the consistency of the full scale, we should consider that the item 'response to eyelid opening' was scored in vivo by a non-blinded observer. This could be considered as a bias (a similar approach was done for the previous EOPS as well (Ortolani et al., 2021) but it couldn't be avoided considering this action as the only possible way to evaluate the pain response around the painful area.

Lastly, we cannot rule out the hypothesis that time-lapse video recording, or simply a longer video recorded from a camera attached to the box wall, could have further improved the Behavioural data scale assessment. The presence of an operator filming with a smartphone might have interfered with the horse's behaviour (Bussières et al., 2008;

Pinho et al., 2020). Pain is a dynamic experience so longer video-clips might have better included different levels of pain in the same video among relevant behaviours as well as changes of facial expression, i.e., the tension of the observed muscles. The real-time application of R-EOPS would pose other challenges but also some advantages. Unlike video, in vivo it is not possible to review the images or zoom in but more details can be noted. Moreover, facial expressions could differ due to the diversity of pain, whether it is acute nociceptive or chronic (Ashley et al., 2005; Hausberger et al., 2016), or depending on other affective states, such as fear and stress (Lundblad et al., 2021). Nevertheless, it is likely that, in a real scenario, the observer's empathy with the horse will never be an irrelevant component in the overall assessment of the animal's welfare, therefore contributing to effective pain management.

Nevertheless, it is necessary to reiterate that the validity is an ongoing process, particularly in this case, where a relatively small, heterogeneous population of horses affected by several ocular diseases has been studied. To confirm all aspects of the validity, a large-scale application and longer-term repetitions would be necessary.

Conclusions

In this study, the validity of a refined version of the previous Equine Ophthalmic Pain Scale was assessed. The use of the new R-EOPS improved objectivity in identifying animals affected by ophthalmic diseases and assessing their pain perception. The exclusion of physiological data increased the consistency of the scale, and we hypothesised that its application would be easier and quicker than the EOPS. The optimal cutoff of 6 showed excellent sensitivity and specificity values, confirming its potential for clinical application. However, adequate training is required, possibly including the use of a booklet where images and related scores are reported, with a particular focus on some Ofex items. Furthermore, a different experimental approach would be useful to evaluate responsiveness over time and to analgesia treatment. Given its easy applicability and the high inter-observer reliability of the TS, this pain scale may be useful in clinical cases for evaluating the degree of pain in horses with ocular or adnexa diseases and for guiding analgesic management as well as the appropriate healing or improvement after treatment. However, there is still room for improvement and validation, as an ongoing process, would require a real-time, large-scale and longer-term application.

Declaration of Competing Interest

None of the Authors has any financial or personal relationship that could inappropriately influence or bias the content of the paper.

Acknowledgements

We would like to thank Dr. Alice Bertoletti, Dr. Samantha Morini, Dr. Matteo Vuerich, Dr. Silvia Napoli, for their help with video analyses and pain scale application.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.tvjl.2024.106079.

References

Ashley, F.H., Waterman-Pearson, A.E., Whay, H.R., 2005. Behavioural assessment of pain in horses and donkeys: application to clinical practice and future studies. Equine Veterinary Journal 37, 565–575.

AWIN, 2015. AWIN Welfare Assessment Protocol for Horse. Doi: 10.13130/AWIN_HORSES_2015.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quiñonez, H.R., Young, S.L., 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Frontiers Public Health 6, 149. https://doi.org/10.3389/fpubh.2018.00149.

Brooks, D.E., Plummer, C.E., 2002. Diseases of the equine cornea. In: Gilger, B.C. (Ed.), Equine Ophthalmology, 4th ed. John Wiley & Sons, Inc, Hoboken, NJ, USA, pp. 253–440.

Bussières, G., Jacques, C., Lainay, O., Beauchamp, G., Leblond, A., Cadoré, J.L., Desmaizières, L.M., Cuvelliez, S.G., Troncy, E., 2008. Development of a composite orthopaedic pain scale in horses. Research in Veterinary Science 85 (2), 294–306.

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences, 2nd ed. Academic Press, New York, NY.

Dai, F., Leach, M., MacRae, A.M., Minero, M., Costa, E.D., 2020. Does thirty-minute standardised training improve the inter-observer reliability of the Horse Grimace Scale (HGS)? A Case Study. Animals 10, 781.

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.

Dalla Costa, E., Stucke, D., Dai, F., Minero, M., Leach, M.C., Lebelt, D., 2016. Using the horse grimace scale (HGS) to assess pain associated with acute laminitis in horses (Equus caballus). Animals 6, 47.

Descovich, K., Wathan, J., Leach, M.C., Buchanan-Smith, H.M., Flecknell, P., Farningham, D., Vick, S.J., 2017. Facial expression: an underutilised tool for the assessment of welfare in mammals. ALTEX 34 (3), 409–429.

Di Giminiani, P., Brierley, V.L., 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. Frontiers in Veterinary Science 3, 100.

Dyson, S.J., Berger, J., Ellis, A., Mullard, J., 2017. Can the presence of musculoskeletal pain be determined from the facial expressions of ridden horses (FEReq)? Journal of Veterinary Behaviour 19, 78–89.

Ekman, P., Friesen, W.V., 1986. A new pan-cultural facial expression of emotion. Motivation and Emotion 10 (2), 159–168.

Field, A., Miles, J., Field, Z., 2009. Discovering Statistics Using SPSS, 3rd ed. SAGE Publications, Ly, London, UK.

Fleiss, J.L., 1986. Reliability of Measurement. In: Wiley, John, Sons, I. (Eds.), The Design and Analysis of Clinical Experiments. Wiley-Interscience Publication, New York, NY, pp. 1–32.

Giammarino, M., Mattiello, S., Battini, M., Quatto, P., Battaglini, L.M., Vieira, A.C.L., Stilwell, G., Renna, M., 2021. Evaluation of inter-observer reliability of animal welfare indicators: which is the best index to use? Animals 11, 1445.

Gleerup, K.B., Andersen, P.H., Munksgaard, L., Forkman, B., 2015a. Pain evaluation in dairy cattle. Applied Animal Behaviour Science 171, 25–32.

Gleerup, K.B., Forkman, B., Lindegaard, C., Andersen, P.H., 2015b. An equine pain face. Veterinary Anaesthesia and Analgesia 42, 103–114.

Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive Veterinary Medicine 45, 23–41.

Gwet, K.L., 2002. Kappa Statistic is Not Satisfactory for Assessing the Extent of Agreement between Raters. Statistical Methods For Inter-Rater Reliability Assessment, No. 1, April 2002.

Gwet, K.L., 2015. On the Krippendorff's Alpha Coefficient. https://www.researchgate.net/publication/267823285_On_Krippendorff's_Alpha_Coefficient.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology 8, 23–34.

Hausberger, M., Fureix, C., Lesimple, C., 2016. Detecting horses'sickness: In search of visible signs. Applied Animal Behaviour Science 175, 41–49.

Knickelbein, K.E., Scherrer, N.M., Lassaline, M., 2018. Corneal sensitivity and tear production in 108 horses with ocular disease. Veterinary Ophthalmology 21, 76–81.

Krippendorff, K., 1970. Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement 30, 61–70.

Langford, D.J., Bailey, A.L., Chanda, M.L., Clarke, S.E., Drummond, T.E., Echols, S., et al., 2010. Coding of facial expressions of pain in the laboratory mouse. Nature Methods 7 (6), 447–449.

Lassaline-Utter, M., Cutler, T.J., Michau, T.M., Nunnery, C.M., 2014. Treatment of nonhealing corneal ulcers in 60 horses with diamond burr debridement (2010-2013). Veterinary Ophthalmology 17, 76–81.

Lundblad, J., Rashid, M., Rhodin, M., Andersen, P.H., 2021. Effect of transportation and social isolation on facial expressions of healthy horses. PLoS ONE 4 (16(6)), e0241532.

Machin, D., Campbell, M.J., Tan, S.B., Tan, S.H., 2009. Sample Size Tables for Clinical Studies. Blackwell Publishing, John Wiley & Sons, Oxford, OX4 2DQ, UK. ISBN: 978-1-4051-4650-0.

Makra, Z., Csereklye, N., Riera, M.M., McMullen Jr, R.J., Veres-Nyéki, K., 2021. Effects of intravenous flunixin meglumine, phenylbutazone, and acupuncture on ocular pain scores in the horse: a pilot study. Journal of Equine Veterinary Science 98, 103375.

Marcantonio Coneglian, M., Duarte Borges, T., Weber, S.H., Godoi Bertagnon, H., Michelotto, P.V., 2020. Use of the horse grimace scale to identify and quantify pain due to dental disorders in horses. Applied Animal Behaviour Science 225, 104970.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica (Zagreb) 22 (3), 276–282.

Meagher, R.K., 2009. Observer ratings: validity and value as a tool for animal welfare research. Applied Animal Behaviour Science 119, 1–14.

Menchetti, L., Righi, C., Guelfi, G., Enas, C., Moscati, L., Mancini, S., et al., 2019. Multi-operator qualitative behavioural assessment for dogs entering the shelter. Applied Animal Behaviour Science 213, 107–116.

Menchetti, L., Dalla Costa, E., Minero, M., Padalino, B., 2021. Development and validation of a test for the classification of horses as broken or unbroken. Animals 11, 2303.

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. Neuroscience & Biobehavioral Reviews 116, 480–493.

Mokkink, L., Terwee, C., de Vet, H., 2021. Key concepts in clinical epidemiology: responsiveness, the longitudinal aspect of validity. Journal of Clinical Epidemiology 140, 159–162.

Oliver, V., De Rantere, D., Ritchie, R., Chisholm, J., Hecker, K.G., Pang, D.S., 2014. Psychometric assessment of the Rat Grimace Scale and development of an analgesic intervention score. PLoS One 9 (5), e97882.

Orth, E.K., Navas González, F.J., Iglesias Pastrana, C., Berger, J.M., Jeune, S.S.L., Davis, E.W., McLean, A.K., 2020. Development of a donkey grimace scale to recognize pain in donkeys (Equus asinus) post castration. Animals 13 (10(8)), 1411.

Ortolani, F., Scilimati, N., Gialletti, R., Menchetti, L., Nannarone, S., 2021. Development and preliminary validation of a pain scale for ophthalmic pain in horses: the Equine Ophthalmic Pain Scale (EOPS). The Veterinary Journal 278 (8), 105774.

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.

Price, J., Catriona, S., Welsh, E.M., Waran, N.K., 2003. Preliminary evaluation of a behaviour-based system for assessment of post-operative pain in horses following arthroscopic surgery. Veterinary Anaesthesia and Analgesia 30, 124–137.

Pritchett, L.C., Ulibarri, C., Roberts, M.C., Schneider, R.K., Sellon, D.C., 2003. Identification of potential physiological and behavioral indicators of postoperative pain in horses after exploratory celiotomy for colic. Applied Animal Behaviour Science 80, 31–43.

Prkachin, K.M., 2009. Assessing pain by facial expression: facial expression as nexus. Pain Research and Management 14, 53–58.

Prucha, V.J.S., Tichy, A., Nell, B., 2020. Equine non-healing corneal ulcers: a retrospective evaluation of 57 cases (2001–2017). Tierarztlich- Praxis Ausgabe Giornale Grosstiere Nutztiere 48, 92–97.

Rankins, E.M., Manso Filho, H.C., Malinowski, K., McKeever, K.H., 2022. Muscular tension as an indicator of acute stress in horses. Physiological Reports 10 (6), e15220.

Richardson, C.A., Flecknell, P.A., 2005. Anaesthesia and post-operative analgesia following experimental surgery in laboratory rodents: are we making progress? Alternatives to Laboratory Animals 33 (2), 119–127.

Riva, M.G., Sobrero, L., Menchetti, L., Minero, M., Padalino, B., Dalla Costa, E., 2022. Unhandled horses classified with broken/unbroken test (BUT) exhibit longer avoidance, flight reactions, and displacement behaviors when approached by humans. Frontiers in Veterinary Science 9, 1022255.

Silva, N.E.O.F., Trindade, P.H.E., Oliveira, A.R., Taffarel, M.O., Moreira, M.A.P., Denadai, R., Rocha, P.B., Luna, S.P.L., 2020. Validation of the Unesp-Botucatu composite scale to assess acute postoperative abdominal pain in sheep (USAPS). PloS One 15, e0239622.

Sotocinal, S.G., Sorge, R.E., Zaloum, A., Tuttle, A.H., Martin, L.J., Wieskopf, J.S., et al., 2011. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Molecular Pain 7, 1744–8069.

Streiner, D.L., Norman, G.R., 2008. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford University Press, Oxford, U.K.

Sutton, G.A., Paltiel, O., Soffer, M., Turner, D., 2013. Validation of two behaviour-based pain scales for horses with acute colic. The Veterinary Journal 197, 646–650.

Taha, A., Darwish, A., Hassanien, A.E., 2017. Arabian Horse Identification Benchmark Dataset. arXiv, 1706.04870.

Tallon, R., Hewetson, M., 2021. Inter-observer variability of two grading systems for equine glandular gastric disease. Equine Veterinary Journal 53 (3), 495–502.

Van Dierendonck, M.C., van Loon, J.P., 2016. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): a validation study. The Veterinary Journal 216, 175–177.

Van Dierendonck, M.C., Burden, F.A., Rickards, K., van Loon, J.P., 2020. Monitoring Acute Pain in Donkeys with the Equine Utrecht University Scale for Donkeys Composite Pain Assessment (EQUUS-DONKEY-COMPASS) and the Equine Utrecht University Scale for Donkey Facial Assessment of Pain (EQUUS-DONKEY-FAP). Animals 10, 354.

van Loon, J.P.A.M., Van Dierendonck, M.C., 2015. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): a scale-construction study. The Veterinary Journal 206, 356–364.

van Loon, J.P.A.M., Van Dierendonck, M.C., 2017. Monitoring equine head-related pain with the Equine Utrecht University scale for facial assessment of pain (EQUUS-FAP). The Veterinary Journal 220, 88–90.

van Loon, J.P.A.M., Van Dierendonck, M.C., 2019. Pain assessment in horses after orthopaedic surgery and with orthopaedic trauma. The Veterinary Journal 246, 85–91.