Validation of a Facial Expression based Pain Scale (FEPS) for objective pain assessment of donkeys with chronic painful diseases.

Veterinary Medicine, Master Research Project

Student: Lyanne Bardelmeijer
Student number: 4183940
Supervisor: J.P.A.M. van Loon
March – June 2019.

Table of contents

Summary

Over the last decades animal welfare has become more important. In order to assess a donkeys welfare, we would like to develop an objective way to assess chronic pain in donkeys. Facial expression based pain scales have been developed for horses. Because of the differences between horses and donkeys, they need some adjustment to be useful for donkeys. The aim of this study is to validate the Facial Expression Pain Scale for donkeys (FEPS$_{donkey}$) for chronic pain, which was already previously investigated for validation for acute pain.

A total of 77 donkeys were observed by 2 observers. The 38 patient donkeys were all previously diagnosed with a chronic painful disease by the veterinarians from the Donkey Sanctuary in England. The 39 healthy control donkeys were matched and selected by the observers. The donkeys were scored twice a day for 3 consecutive days. Furthermore a videotape of the donkeys faces during locomotion was made in order to assess facial expressions during locomotion. A survey was developed and filled out by the grooms to get a better impression of the donkeys' overall health.

A high interobserver agreement was found between observer 1 and 2 (Intraclass Correlation Coefficient (ICC) = 0,96 with p-value <0,001). There was a significant difference between patient donkey scores and control donkey scores with the FEPS$_{donkey}$ (p<0,001). Overall sensitivity of 80,6% and specificity of 56,3% was found. There was no correlation between the FEPS$_{donkey}$ score and the survey score and a weak correlation between the FEPS$_{donkey}$ score and the Visual Analogue Scale (VAS) score (ICC = 0,56 with p<0,01). There was a good inter- and intra-observer reliability for scoring the videos with FEPSdonkey (ICC respectively 0,69 and 0,76 with p-value <0,001), however no significant difference was found between patients and controls (p-value = 0,19).

FEPS$_{donkey}$ proved to be reliable and clinically applicable to use for recognising chronic pain in donkeys. A larger amount of donkeys used in a future study is needed to further validate pain assessments. The survey and the FEPS$_{donkey}$ scoring during locomotion did not prove to be clinically applicable in this set-up. Further research is needed to develop a survey that can aid in assessment of donkeys with possible chronic pain.

Introduction

Background of the study

Over the last decades animal welfare has become more important. The Farm Animal Welfare Committee has developed five freedoms in order to assess animal's welfare, see Fig. 1 (Webster, 2001). These freedoms comprise of being free from hunger and thirst but also of pain, injury and disease. Assessing if an animal is free from pain can be a difficult task. Visual analogue scales can be used to question owners of where they think their animal is on a scale of no pain to maximum pain imaginable. Unfortunately, the inter-observer reliability for the VAS score is only fair if it needs to be used to assess pain in another person or animal, meaning the pain score is not very reliable (Lindegaard et al., 2010) .

The Five Freedoms and Provisions

1. *Freedom from thirst, hunger and malnutrition* – by ready access to fresh water and a diet to maintain full health and vigour.
2. *Freedom from discomfort* – by providing a suitable environment including shelter and a comfortable resting area
3. *Freedom from pain, injury and disease* – by prevention or rapid diagnosis and treatment.
4. *Freedom to express normal behaviour* – by providing sufficient space, proper facilities and company of the animal's own kind.
5. *Freedom from fear and distress* – by ensuring conditions which avoid mental suffering.

Figure 1: the five freedoms and provisions in order to assess animal's welfare

The Composite Pain Score

A more objective pain score can improve veterinarians' ability to recognise the pain signals the animal is expressing and take adequate intervention steps. These pain scores often contain multiple behavioural expressions and physiological expressions that can change when an animal is in pain. Pain scores that comprise different behavioural and physiological elements are called composite pain scales (CPS) (Sutton et al., 2013, Bussieres et al., 2008). For horses different pain scales have been described for acute pain combining the various behavioural changes seen with palpating a painful area, posture, pawing on the floor, kicking at the abdomen and head movement. An example for behavioural pain scales is the Equine Acute Abdominal Pain Scales (EAAPS), assessing the changes in behaviour (EAAPS-1) and the frequency of the behaviour being demonstrated (EAAPS-2) for horses in acute pain (Sutton, Gila Abells et al., 2013, Sutton et al., 2013). Physiological parameters as respiratory rate, heart rate and temperature can also be included in a composite pain scale. Based on physiological parameters, behavioural expression changes and the frequency of this behaviour being demonstrated, the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) has been developed and validated for acute visceral pain in horses (van Loon, Van Dierendonck., 2015, Van Dierendonck, van Loon., 2016) .

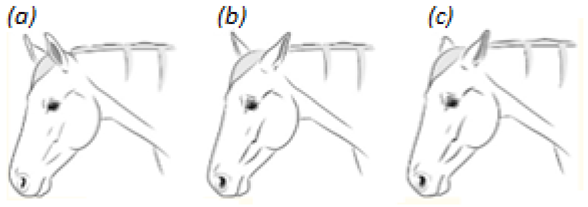


Figure 2: (a) Facial expression of a pain free, relaxed and attentive horse (Ill. Andrea Klintbjer). (b) Facial expression of a horse in pain, comprising all features of the pain face including asymmetrical ears (Ill. Andrea Klintbjer). (c) Facial expression of a horse in pain, comprising all features of the pain face including low ears (Ill. Andrea Klintbjer).

The Facial Expression Pain Score

Another strategy for scoring and recognising pain is looking at the changes in facial expressions during pain. For horses, changes in ear positioning, a tense or withdrawn stare, dilated nostrils, tension of the lips and other facial muscles have been described as an equine pain face, see Fig. 2 (Gleerup et al., 2015). The Horse Grimace Scale (HGS) has been developed to assess differences in facial expressions before and after surgery. The horses were being scored by looking at facial action units such as ear positioning, orbital tightening, tension above the eyes, straining of the chewing muscles, straining of the mouth and pronounced chin, straining of the nostrils and flattening of the profile (Dalla Costa, 2014). Another scale for scoring facial expression changes during pain is the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP). This has been developed and validated for acute visceral pain such as colic, based on the amount of facial expression changes that were seen during a two minute observation (Van Dierendonck, van Loon., 2016, van Loon, Van Dierendonck., 2015). The EQUUS-FAP has also been validated for acute head related pain such as sinusitis, alveolitis, uveitis, mandibular or maxillary fractures and post-operative pain after head related surgery (van Loon, Van Dierendonck, 2017). Recently the EQUUS-FAP was validated for horses with acute orthopaedic trauma and for horses after orthopaedic surgery (van Loon, Van Dierendonck, 2019).

Adjusting scales from horses to donkeys

The EQUUS-COMPASS and EQUUS-FAP have been developed for assessment of specific types of acute pain in horses. For donkeys these scales needed some adjustments and therefore, new studies for validation of these adjusted pain scales. This is necessary because donkeys have been reported to express their pain differently than horses (Burden, Thiemann, 2015). Compared to the horse, a donkey is less of a flight- and more of a fight animal. Donkeys mask their signs of pain by being stoic, resulting in more subtle behavioural and facial expression changes (Ashley, Waterman-Pearson & Whay, 2005). Based on this, the EQUUS-COMPASS and EQUUS-FAP have been adjusted for donkeys, resulting in the Donkey Composite Pain scale (Do-CPS) and the facial expression pain scale donkey ($FEPS_{donkey}$) (Van Dierendonck et al., 2018), see Fig. 3. A large cohort of patients has been studied with these scales for further validation (study under review).

Adjusting scales from acute pain to chronic pain

Pain scores for donkeys with chronic pain have not been developed yet. A recent master research project at Utrecht University claimed acute orthopaedic pain in horses is being expressed differently than chronic orthopaedic pain (Enström, 2018), indicating that chronic orthopaedic pain needs pain scales with different parameters, compared to those for acute pain. Prior to this research project, the Do-CPS has been adjusted for chronic pain in a pilot study, resulting in the Donkey Chronic Composite pain scale (DO-CCPS). The $FEPS_{donkey}$ has not been adjusted for chronic pain. This study was set up for assessing validity and clinical applicability of the $FEPS_{donkey}$ specifically for chronic pain.

Research goals

The objective of this study was to 1) adjust the FEPSdonkey for acute pain to chronic pain, to 2) analyse the validity and applicability of the FEPSdonkey for chronic pain, to 3) assess if facial expression characteristics can be observed during locomotion in donkeys with the FEPSdonkey and to 4) create a survey for the grooms or caretakers of the donkeys that can aid in assessment of chronic pain in donkeys.

Hypothesis

H0 = the FEPSdonkey is not reliable and applicable for analysing chronic pain in donkeys.
H1 = the FEPSdonkey is reliable and applicable for analysing chronic pain in donkeys.

Materials and methods

The Facial Expression Pain Scale for Donkeys

Facial Expression	Description	Score
Head	Normal movement Less/no or more/ exaggerated movement	0 2
Eyelids	Opened More opened eyes or tightening of eyelids Obviously more opened eyes or obvious tightening of eyelids	0 1 2
Focus	Focused on environment Less focused on environment Not focused on environment	0 1 2
Nostrils	Relaxed A bit more opened, nostrils lifted, wrinkles seen Obviously more opened, nostril flaring, possibly audible breathing	0 1 2
Corners mouth/lips	Relaxed Lifted	0 2
Muscle tone head	No fasciculation's Mild fasciculation's Obvious fasciculation's	0 1 2
Flehming/yawning/smacking	Not seen Seen	0 2
Teeth grinding and/or moaning	Not been heard Heard	0 2
Ear response	Clear response with both ears or ear closest to source Delayed/reduced response to sounds No response to sounds	0 1 2
Ear position	Normal position Abnormal position (hang down/backwards)	0 2
Startle/headshaking	No startle/headshaking At least one startle (a sudden abrupt movement with the head as if suddenly aware of danger)/period of head shaking	0 2
Sweating behind the ears	No signs of sweating Signs of sweating	0 2
Total		/24

Figure 3: The Facial Expression Pain Scale for Donkeys, used to assess chronic pain.

Adjusting FEPSdonkey for chronic pain

In the previously mentioned pilot study the Do-CPS was adjusted for chronic pain, resulting in the Do-CCPS. This pilot study also assessed the need for adjusting the FEPSdonkey previously used for acute pain, but did not find any adjustments necessary for chronic pain scoring. In order to be able to compare the data, it was decided to not adjust the FEPSdonkey any further for chronic pain. The pain score sheet that was used in this study is shown in Fig. 3.

Animals

The Donkey Sanctuary in Devon, England, provided the donkeys used for pain scoring. The veterinarians at the Donkey Sanctuary all have their own farm to look after and thus are familiar with the donkeys' clinical histories on their own farm. Each veterinarian provided a list of donkeys previously diagnosed with chronic painful diseases, varying from chronic osteoarthritis, chronic laminitis, chronic dental disease and other chronic painful diseases. The

Donkey Sanctuary uses an online Animal Management System (AMS) to keep their clinical notes and diagnoses, therefore AMS was used to look up more information about the specific donkeys such as their weight, age and medication management. For each patient donkey, one healthy control donkey was selected. The control donkey was preferably selected from the same group as the patient donkey and needed to be the same gender and approximately the same age and weight as the patient donkey. In case there was no good match for a patient donkey in the same group, a control donkey from a different group, barn or farm was selected. Initially 38 patient donkeys and 43 control donkeys were scored using FEPS$_{donkey}$. After re-assessment with the farm veterinarian, 4 control donkeys were excluded from the study. This was due to an acute onset of muscular discomfort in the back of one donkey, chronic weight loss of one donkey, behavioural problems of another donkey and recurring colitis of the last donkey. One control donkey turned out to have thin soles and was transferred to the patient group. One patient donkey was thought to have chronic laminitis, but due to miscommunication turned out to have had one acute laminitic episode two months prior. Therefore, he was transferred to the control group. In total there were 38 patient donkeys and 39 control donkeys used in this study, see table 1. The donkeys were visually divided into the height categories small, average and large. Mean weight and standard deviation of the small, average and large donkeys was respectively 102,3 (±13,8) kg, 172,7 (±21,8) kg and 247,4 (±38,1) kg.

Table 1: 38 patient donkeys and 39 control donkeys were used in this study. The patient donkeys were divided into multiple categories based on their clinical diagnosis. For each group the amount of geldings versus mares is listed, the mean age, the breed, the mean weight and whether or not they received any pain relief medication. No stallions were observed in this study.

	Chronic condition	Number of donkeys	Sex: gelding	Sex: mare	Mean age in years (±SD)	Breed small	Breed average	Breed large	Mean weight in kg (±SD)	On pain relief medication	No medi-cation
Patients	Osteoarthritis	21*	13	8	21,3 (±7,2)	4	11	6	185,5 (±63,7)	19	2
	Chronic Laminitis	10	2	8	21,4 (±5,6)	1	6	3	190,5 (±42,3)	8	2
	Chronic Dental disease	10	4	6	23,8 (±6,4)	0	8	2	178,6 (±41,4)	6	4
	Other chronic painful problems	2	1	1	18,5 (±4,9)	0	2	0	159,0 (±7,1)	2	0
Controls		39	22	17	19,1 (±6,1)	0	32	7	185,7 (±37,0)	1	38
Total		77	40	37	20,3 (±6,4)	4	56	17	185,5 (±44,7)	32	45

** Of the osteoarthritic group there were 5 patients that had multiple problems. They were also added to other groups. 3 donkeys from the osteoarthritis group also had chronic dental problems, 1 donkey also had chronic laminitis and 1 donkey also had other chronic painful problems.*

Procedure of scoring

Data collection for validating FEPSdonkey and Do-CCPS was done simultaneously. The observations were done by two veterinary students from the University of Utrecht. The observers made themselves familiar with the FEPS and Do-CCPS scoring system by training with the donkeys from the Donkey Sanctuary in Zeist. Once the observers found themselves comfortable with the scoring system, they started collecting data in England. The observers first scored the donkeys simultaneously for FEPS and then for Do-CCPS and did not discuss their findings during the observation time. The donkeys were scored in their own group and within their normal daily routine.

Each donkey was observed using the FEPSdonkey scoring system and the Do-CCPS twice a day for three consecutive days, except Saturdays and Sundays. The morning assessments started at 8am and ran until 11.30am. Afternoon assessments started at 12.30pm and ran until 4pm. The time periods for the assessments were called

- T=0 (first day, morning),
- T=1b (first day, afternoon),
- T=2a (second day, morning),
- T=2b (second day, afternoon),
- T=3a (third day, morning) and
- T=3b (third day, afternoon).

Once the donkey was identified, the two observers put on a headcollar when necessary and started their observation. In case the donkey was eating, the observers waited for 5 minutes for the donkey to chew and swallow the last bits of food. The donkey was observed for two minutes to determine the score for the FEPSdonkey categories, see Fig. 3. Dynamic categories such as ear position and corners of the mouth were scored abnormal if the donkey showed abnormalities for more than 50% of the observed time. The donkey was observed for five minutes for the Do-CCPS. Preliminary results can be found in the master research report that analyses the Do-CCPS data (Vos, Unpublished results).

Due to limited time in the last week of collecting data, 3 patient donkeys were not observed at time period T=3b, 1 patient donkey was not observed during the T=2b, T=3a and T=3b time periods, 1 control donkey was not observed at time period T=3b and 3 control donkeys were not observed during the T=2b, T=3a and T=3b time periods. This means that a total of 446 scores were collected by each observer, of which 222 belonged to patient donkeys and 224 belonged to control donkeys.

FEPSdonkey during locomotion

In order to determine whether facial expression changes could be scored using FEPSdonkey during locomotion, the observers made a short videotape of the donkey's face while the donkey was walking. After each scoring session with the FEPSdonkey at rest, the donkey was asked to walk while one of the observers recorded the videotape. These videotapes were later edited to short videos of 10 seconds. The videos were randomised, blinded and later shown to two other observers (2 veterinarians from Utrecht University), them being two veterinarians from the University of Utrecht who were also familiar with the FEPS scoring system. The observers were not familiar with the donkeys on the videos. They scored the videos simultaneously and did not discuss their findings during the observation time. A smaller selection of all the videos that were recorded were assessed. It was randomly chosen that that would be the videos taken at day 3 during the morning assessment (T=3a). In case there was no video of that time period, or the video was not useful, a morning assessment video of day

2 (T=2a) or day 1 (T=0) was selected. The observers scored a total of 86 videos, of which 71 were original videos and 15 were doubled to later calculate the intra-observer reliability. An encryption key was developed to determine which score was given to which donkey. Of the 71 original videos, 36 belonged to patient donkeys and 35 belonged to control donkeys. The 15 doubled videos were of 7 patient donkeys and 8 control donkeys, meaning there were 43 videos of patient donkeys and 43 videos of control donkeys in total.

Survey

A survey was developed to be filled out by the grooms or caretakers of the donkeys, in order to get an impression of the donkey's overall health. The questions focused on changes to be seen over time such as changes in the donkey's mood, losing weight and getting reluctant to move. The questions were developed like statements and could be answered by a 6-point scale varying from completely agree to completely disagree, see Fig. 4. The last question on the survey was to ask the owner to give a number to the amount of pain they thought the donkey was in, on a Visual Analog Scale from 0 to 10. The surveys were given to the grooms and filled out in the time period the observers were at the farm to observe the donkeys.

<u>Survey for the owner/caretaker of the donkey</u>

Name stable
Name/number donkey

Please fill in for each statement if this is the case for this donkey.
Please cross the corresponding box.

	Completely disagree					**Completely agree**
The donkeys mind set is alert in general.	o	o	o	o	o	o
The donkey has shown weight loss over time during the last weeks/months.	o	o	o	o	o	o
The donkey interacts less with other donkeys.	o	o	o	o	o	o
The donkey lays down more than he/she used to.	o	o	o	o	o	o
In the case of chronic/returning lameness; the lameness seen today is normal for this donkey. If disagree: please explain:	o	o	o	o	o	o
In the case of chronic dental problems; the donkeys chewing pattern is normal today. If disagree: please explain:	o	o	o	o	o	o

On a scale from 0 to 10, how much pain do you think the donkey is in today?
Please surround the corresponding number.

0 1 2 3 4 5 6 7 8 9 10

Figure 4: the survey used for questioning the groom/caretaker of the donkey, to get an impression of his/her overall health.

Clinical cases

Two examples of clinical cases were included in this report. One case was not included in the data set as a patient donkey, because he was only observed once. The other case was observed six times and is included in the data set. The clinical cases were included because of the relatively high FEPSdonkey and Do-CCPS scores. Both cases were followed up with radiographs and a clinical exam and were later euthanised and send in for post-mortem observations. Radiographs and post-mortem photos were added in the report.

Data processing and statistical analysis

The data was processed using Microsoft Excel 2016, by making box-and-whisker plots which show the minimum, maximum and mean scores, as well as the quartiles and medians. A cut-off value was determined by trying to obtain maximum differentiation between patient donkey scores and control donkey scores. The determined cut-off value was used to calculate the sensitivity and specificity of the total $FEPS_{donkey}$ scoring system. For each individual scoring parameter the sensitivity and specificity was calculated as well, using a cut-off value of >0. Based on the sensitivity and specificity, some parameters were then excluded from scoring system and overall sensitivity and specificity was calculated again. Sensitivity and specificity was also calculated for all the individual survey questions, in which a cut-off value of >0 was used.

The statistical analysis of the data was done by using IBM SPSS Statistics Version 25. The inter-observer reliability was calculated by using the Intraclass Correlation Coefficient, determining the value of Cronbach's alpha and the p-value. As the inter-observer reliability was high, the pain scores of the two observers were combined to an average pain score. The Friedman test was then used to determine if there was a significant difference over the six observation times of respectively patient donkeys and control donkeys. The Friedman test showed no significant difference over the six time periods for the patient donkeys, see supplementary data 1. The patient donkeys were further divided into groups representing their medication management and the Friedman test was done to determine if there was any significant difference between morning and afternoon scores for different medication managements. The Friedman test showed no significant difference over time for the individual medication management groups, see supplementary data 2. This meant that the patient donkey scores could be averaged into one total score. However, when the Friedman test was used to determine significant differences over time for the control donkeys, this did show a significant difference, see supplementary data 3. This meant that the scores of the control donkeys could not be averaged to one total score. Therefore, for the individual day scores the morning and afternoon scores were compared. The Wilcoxon test was performed to see if there was a significant difference between the individual morning and afternoon assessments per day, see supplementary data 3. The test showed no difference in morning or afternoon assessments for each day for respectively patient and control donkeys, so three day averages for each donkey were calculated. Then, the Friedman test was performed again to see if there was still a significant difference between the three average day scores for respectively patient and control donkeys. The test found no difference for patient donkeys, but again did find a significant difference for control donkeys between the day averages. This meant the day average scores could not be further averaged into a total score, however in order to compare patient donkey scores to control donkey scores, these day average scores could be used. To see if there was a significant difference between patient donkey scores and control donkey scores, and therefore to see if the $FEPS_{donkey}$ scoring system was applicable, the patient donkey day averages and control donkey day averages for day one, two and three were compared using the Mann-Whitney U test.

For the statistical analysis of the $FEPS_{donkey}$ scores during locomotion, the inter-observer reliability was determined by using the Intraclass Correlation Coefficient to calculate the value of Cronbach's alpha and the p-value. As the inter-observer reliability was high, the pain scores of the two observers were combined to an average pain score. The Mann-Whitney U test was then used to determine if there was a significant difference between the score of patient donkeys and control donkeys. This test was also done for the specific orthopaedic patient group and the specific dental patient group versus control donkeys. To determine the intra-observer reliability, the two observers scored 15 videos twice. All 30 doubled scores were used to calculate the intra-observer reliability by using the Intraclass Correlation Coefficient.

For the statistical analysis of the survey scores, a total score was calculated by summing up the scores from the individual 6-point scale questions. When calculating the total score, it seemed possible that one farm had interpreted the questions in the survey in a different way than they were intended and had therefore scored them differently, see supplementary data 4 for all the survey results. The survey results from that particular farm, 5 patient donkeys and 5 control donkeys, were excluded from the statistical analysis. As the survey was only filled out for one day and no total average scores could be made, in order to determine any correlation to the FEPSdonkey score, this could only be done against one average dayscore. The correlation of the survey total score was compared to the randomly chosen FEPSdonkey average day 1 score by using the Intraclass Correlation Coefficient. The same test was then used to compare the survey total score to the VAS score and then again to compare the VAS score to the FEPSdonkey average day 1 score. Furthermore, the correlation of each individual survey question against the FEPSdonkey average day 1 score was determined by using the Intraclass Correlation Coefficient, to see which questions had the best correlation to the FEPSdonkey score.
Then, the total survey score from the patient donkeys was compared to the score of the control donkeys by using the Mann Whitney U test. The same test was used to compare the VAS score from the patient donkeys to the control donkeys. For all tests statistical significance was accepted at p-value <0,05.

Results

Interobserver reliability

In Fig. 5 a scatterplot of the interobserver reliability is shown. A Cronbach's Alpha of 0,96 was found with p<0,001.

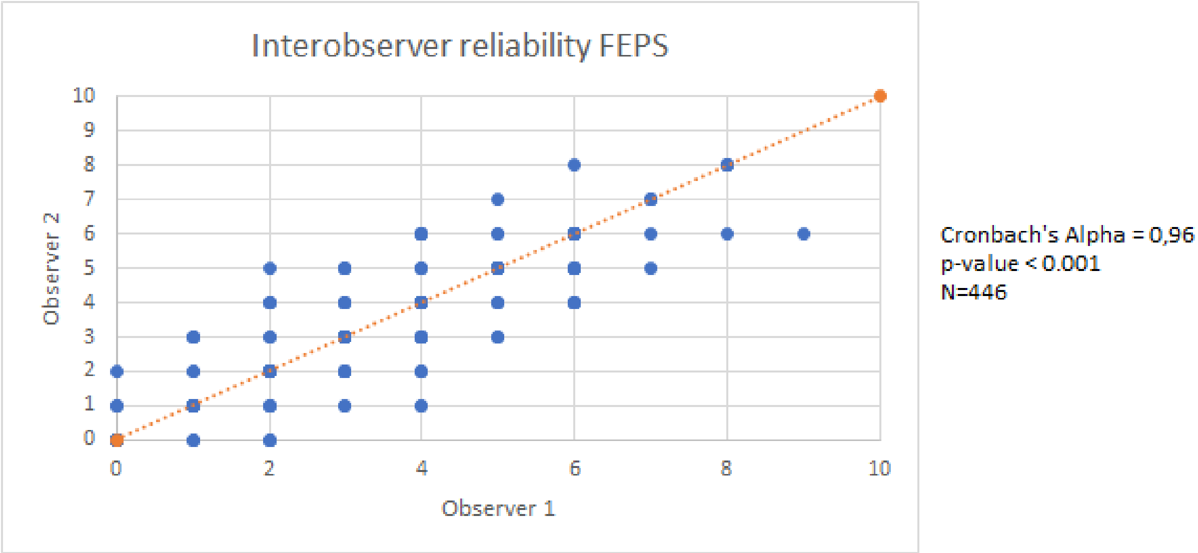


Figure 5: The interobserver reliability of FEPSdonkey in a scatterplot. A total of 446 scores from observer one were set out against the 446 scores from observer two. The red trendline represents the perfect agreement between the two observers. A Cronbach's Alpha of 0,96 was found with p-value <0,001 and N=446.

Differences between the time periods

For the patient donkeys, no significant difference was found between the day averages of day 1, 2 and 3 with p-value = 0,18. Control donkeys showed a significant difference over the three days, with p<0,001, see Fig. 6. Therefore, the scores could not be further averaged and in order to compare patient donkey scores and control donkey scores, the separate average day scores needed to be used.

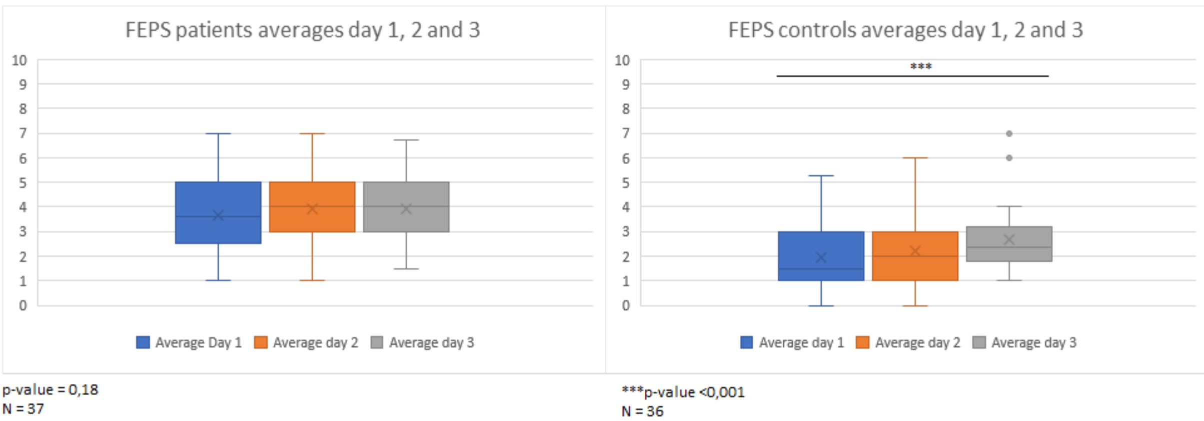


*Figure 6: The calculated day averages for the patient donkeys (left) showed no significant difference between day 1, 2 and 3, with p-value = 0,18 and N = 37. The calculated day averages for control donkeys (right) showed a significant difference between day 1, 2 and 3 with ***p<0,001 and N=36. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners .*

Pain scores of patient donkeys versus control donkeys

The day averages were compared to the control day averages in order to determine the applicability of the FEPSdonkey scoring system. There was a significant difference between patient donkey scores and control donkey scores, with p<0,001 for all individual days, see Fig. 7.

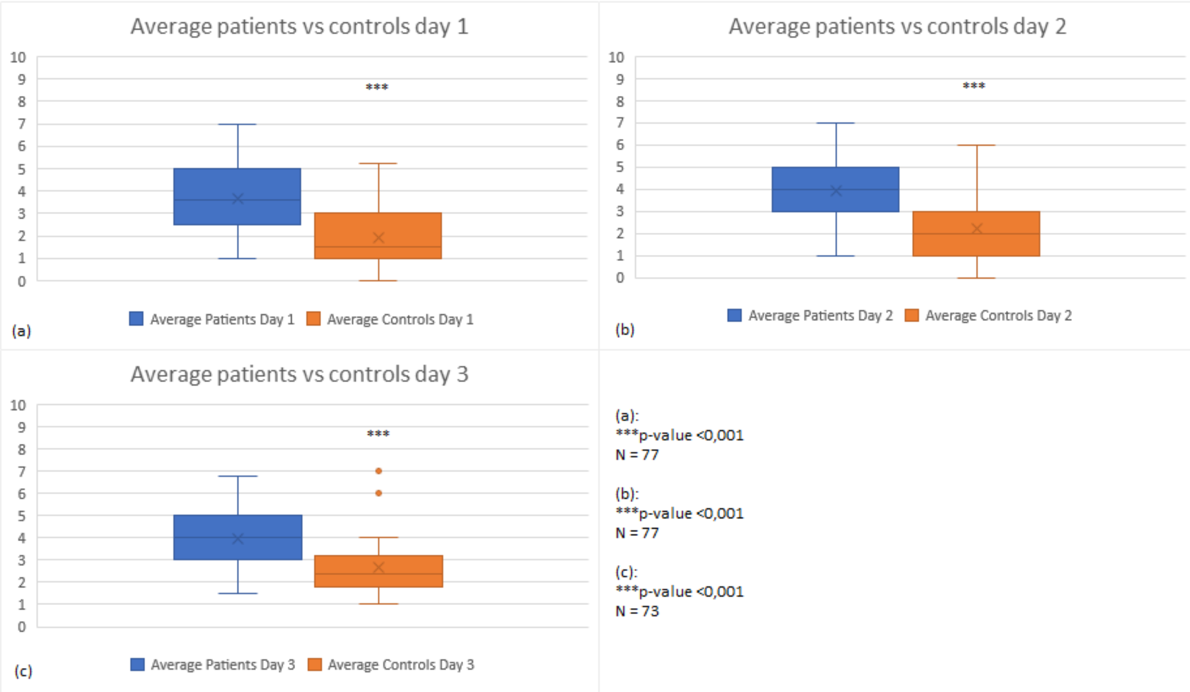


*Figure 7: All days showed significant difference with ***p<0,001 and N=77 (a, b) and N=73 (c). Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.*

Sensitivity and specificity of FEPSdonkey

FEPSdonkey was calculated to have an overall sensitivity of 80,6% and specificity of 56,3%. Furthermore the sensitivity and specificity of each parameter individually were calculated. Table 2 shows the individual sensitivity and specificity for each parameter in percentages.

Table 2: The sensitivity and specificity for each individual scoring parameter of FEPSdonkey, based on cut-off value of >0. Total sensitivity and specificity was based on cut-off value of >2.

Parameter	Sensitivity (%)	Specificity (%)
Head	0	100
Eyelids	65,1	75,7
Focus	0	100
Nostrils	66,0	55,6
Corners mouth/lips	33,2	88,8
Muscle tone head	0,2	99,6
Flehming/yawning/smacking	33,8	70,8
Teeth grinding and/or moaning	0	99,8
Ear response	0	100
Ear position	57,9	62,9
Startle/headshaking	0,5	100

Sweating behind the ears	0	100
Total	80,6	56,3

Based on the individual sensitivity and specificity, there are several parameters that were rarely or never seen: head, focus, muscle tone head, teeth grinding and/or moaning, ear response, startle/headshaking and sweating behind the ears. These parameters all showed a sensitivity close to 0% and a specificity close to 100%. When these parameters were excluded from the data, the overall sensitivity and specificity was calculated again. Sensitivity did not change, but specificity improved with 0,04% from 56,3% to 56,7% when the parameter muscle tone of the head was excluded (Table 3).

Table 3: Sensitivity and specificity of FEPSdonkey including all parameters and after excluding several parameters, which all previously showed a low sensitivity (0-0,5%) and a high specificity (99,6-100%).

	Sensitivity (%)	**Specificity (%)**
FEPS including all 12 parameters	80,6	56,3
FEPS after elimination of the 6 parameters - Head - Focus - Teeth grinding and/or moaning - Ear response - Startle/headshaking - Sweating behind the ears	80,6	56,3
FEPS after elimination of the 7th parameter - Muscle tone head	80,6	56,7

Survey results

Fig. 8. shows the box-and-whiskers plot of respectively the total survey scores and VAS scores of the patient donkeys versus control donkeys. A significant difference was found between the total survey scores of patient donkeys and control donkeys with p-value <0,05. A significant difference was also found between the VAS scores of patient donkeys and control donkeys with p-value <0,001.

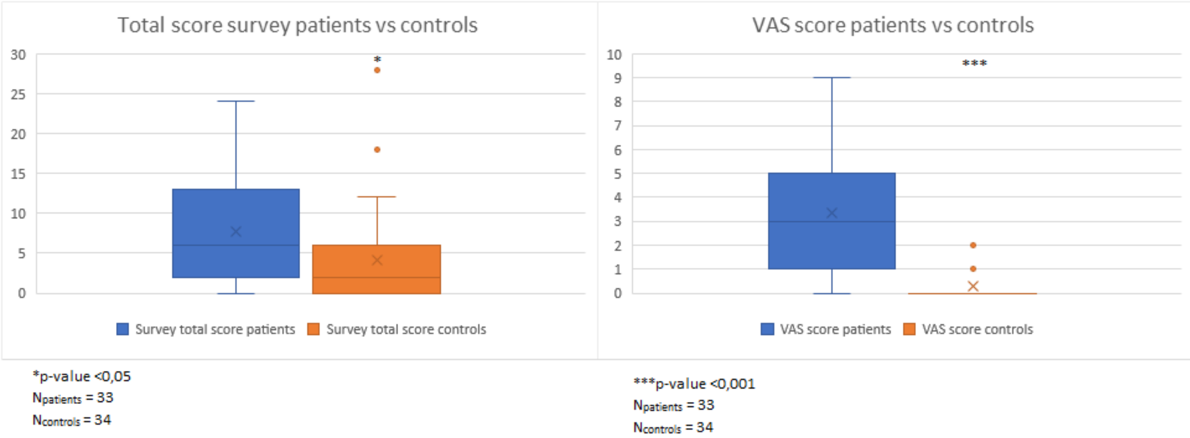


*Figure 8: Total survey scores and VAS scores of patients were significantly higher compared to the scores of control donkeys. Respectively a *p-value of <0,05 and ***p-value of <0,001 was found with $N_{patients}$ = 33 and $N_{controls}$ = 34. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.*

In order to see which survey questions could contribute the most, the sensitivity and specificity for the individual survey questions were calculated (Table 4). Sensitivity and specificity for the specific orthopaedic and dental questions were calculated for that specific patient group.

Table 4: Sensitivity and specificity for the survey questions. Survey questions consisted of a 6-point scale. For the calculations 0 was counted as negative and 2, 4, 6, 8 and 10 were counted as positive. Sensitivity and specificity for the specific orthopaedic and dental questions were calculated for the specific patient groups.
**N = 26. **N = 9.*

Question in the survey	**Sensitivity (%)**	**Specificity (%)**
Mindset is alert	27,2	85,3
Donkey has shown weight loss	60,6	61,8
Donkey interacts less	45,5	70,6
Donkey lays down more	15,2	94,1
In case of orthopaedic problems: lameness is normal	23,1*	88,2
In case of dental problems: chewing pattern is normal	33,3**	94,1

The scatterplot in Fig. 9 shows no correlation between the average day 1 $FEPS_{donkey}$ score and the survey total score with Cronbach's alpha of 0,23 and p-value = 0,15. There is also no correlation between the survey total score and the VAS score with Cronbach's alpha of 0,28 and p-value = 0,09. A weak correlation was determined between the average day 1 FEPSdonkey score and the VAS score with Cronbach's alpha of 0,56 and p-value <0,01.

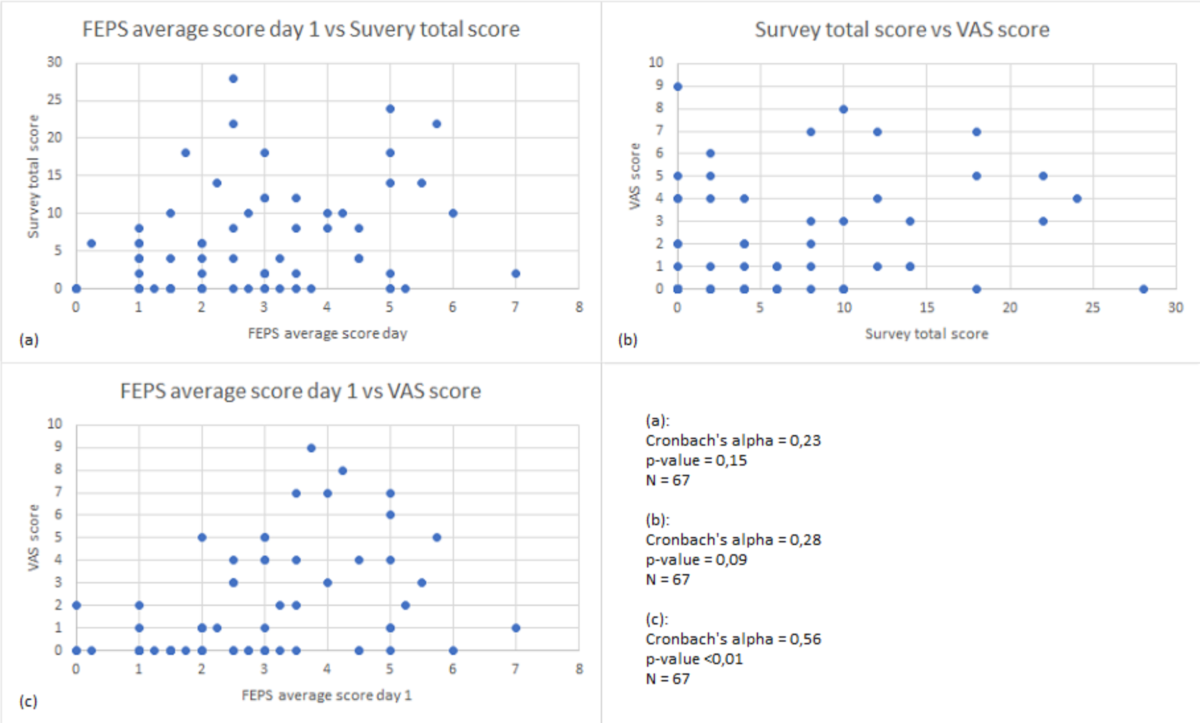


Figure 9: Scatterplot of correlation between the survey score, VAS score and FEPS average day 1 score. Correlation determined for the Survey total score versus VAS score, FEPSdonkey day 1 average score versus VAS score and FEPSdonkey day 1 average score versus Survey total score. (a) FEPSdonkey average day 1 score versus Survey total score shows a Cronbach's alpha of 0,23 with p-value = 0,15 and N = 67. (b) Survey total score versus VAS score shows a Cronbach's alpha of 0,28 with p-value = 0,09 and N = 67. (c) FEPSdonkey average day 1 score versus VAS score shows a Cronbach's alpha of 0,56 with p-value <0,01 and N = 67.

Only the survey question 'The donkey interacts less with other donkeys' had a weak correlation to the FEPSdonkey average day 1 score with a Cronbach's alpha of 0,41 and p-value <0,05 (Table 5). The other questions showed no correlation between the survey question and the FEPSdonkey average day 1 score.

Table 5: The correlation of each individual survey question to the FEPSdonkey average day 1 score was determined.

	Cronbach's alpha	**p-value**	**N**
Mindset is alert	0,30	0,07	67
Donkey has shown weight loss	0,24	0,14	67
Donkey interacts less	0,41	0,02	67
Donkey lays down more	0,30	0,08	67
In case of orthopaedic problems: lameness is normal	0,28	0,10	60
In case of dental problems: chewing pattern is normal	0,09	0,38	43
Total survey score	0,23	0,15	67

FEPSdonkey scoring during locomotion

In Fig. 10 the scatterplot of the interobserver reliability of the FEPSdonkey scoring system during locomotion is shown. A moderate interobserver reliability was found with a Cronbach's alpha of 0,69 was found, with p-value <0,001.

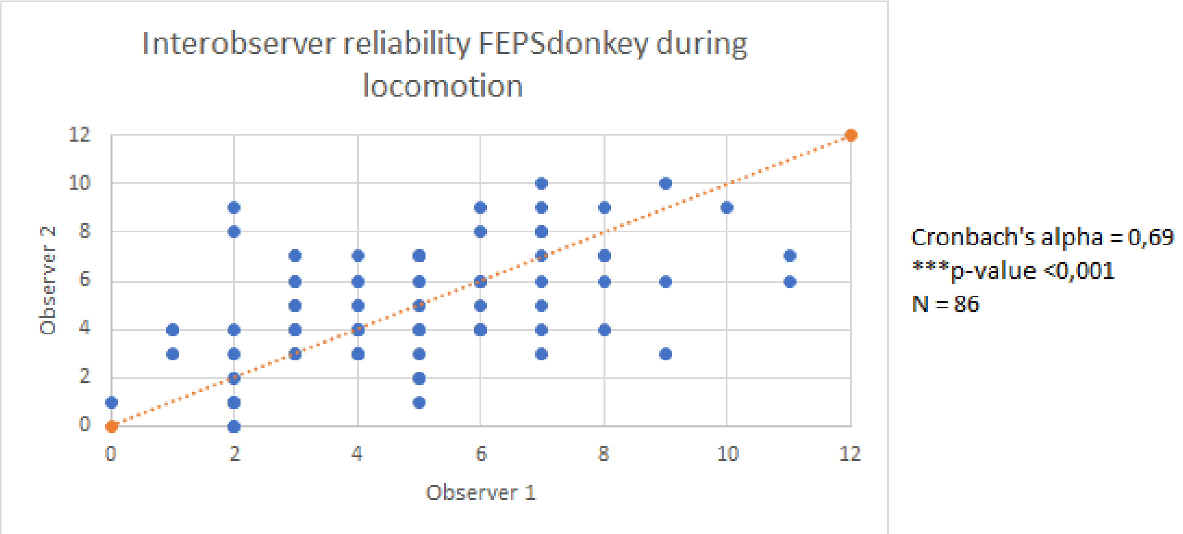


Figure 10: Scatterplot of the interobserver reliability of FEPSdonkey scoring during locomotion. A moderate reliability was found, with a Cronbach's alpha of 0,69, with p-value <0,001 and N =86.The red trendline represents the perfect agreement between observer 1 and 2.

The scatterplot for the intra-observer reliability can be seen in Fig 11. A Cronbach's alpha of 0,76 was found, with p-value <0,001.

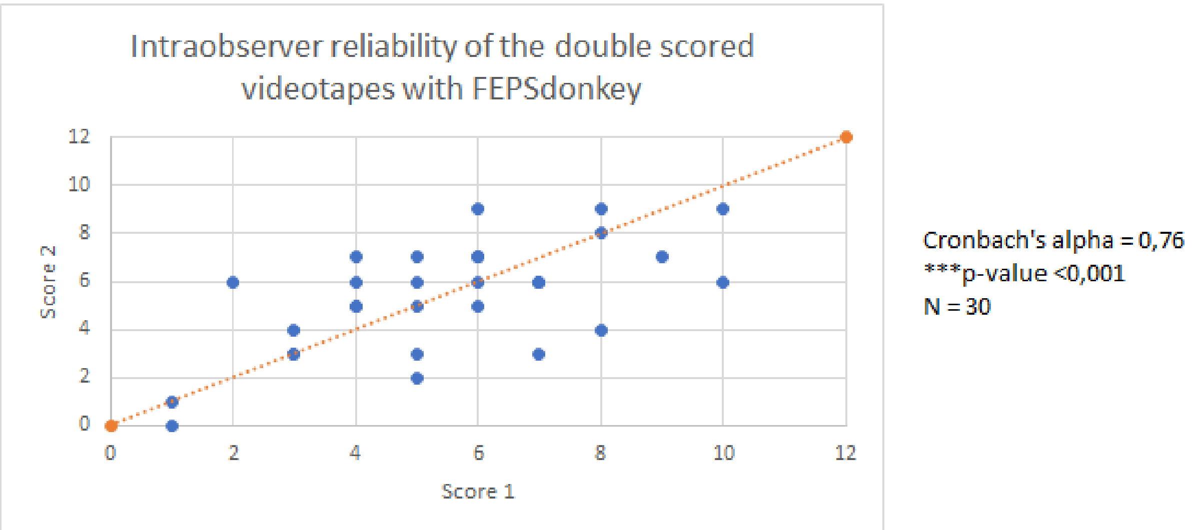


Figure 11: Scatterplot of the intraobserver reliability of FEPSdonkey scoring during locomotion. A good correlation between the first score and second score was found with Cronbach's alpha of 0,76 with p-value <0,001 and N = 30. The red trendline represents the perfect agreement between score 1 and 2.

Furthermore, no differences in pain scores during locomotion were found between patients and control donkeys, with p-value = 0,19, see Fig. 12.

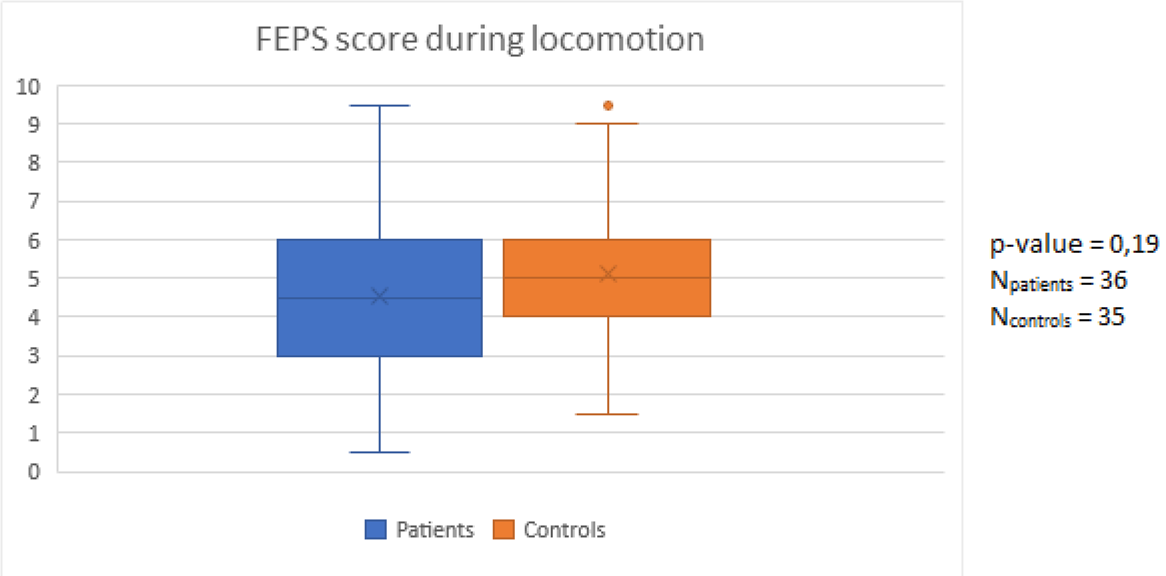
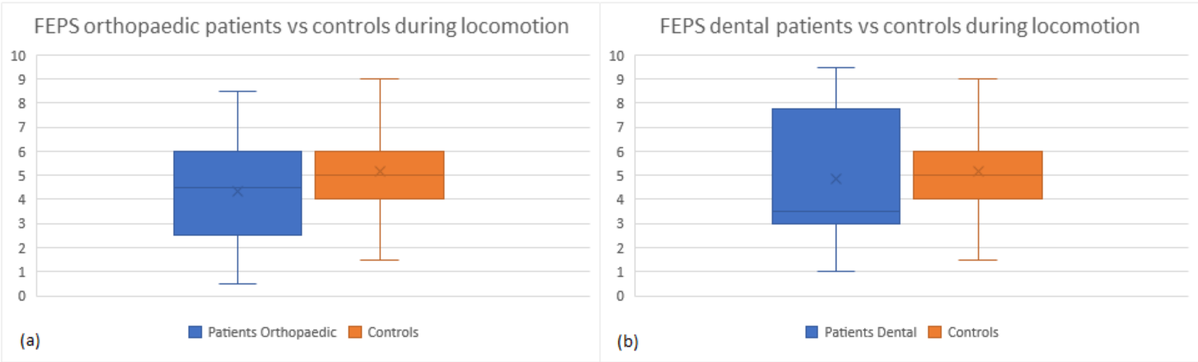


Figure 12: Comparison of FEPSdonkey score during locomotion of patient donkeys and control donkeys. No significant difference between patient scores and control scores was found, with p-value = 0,19 and $N_{patients}$ = 39 and $N_{controls}$ = 35. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.

When selecting only the orthopaedic patient group and dental patient group versus control donkeys, both showed no significant difference between patient donkey scores and control donkey scores for pain assessment during locomotion, see Fig. 13, with p-values of respectively 0,10 and 0,39.



(a):
p-value = 0,10
$N_{patients}$ = 31
$N_{controls}$ = 35

(b):
p-value = 0,39
$N_{patients}$ = 9
$N_{controls}$ = 35

Figure 13: Comparison of FEPSdonkey score during locomotion of specific orthopaedic patient group and dental patient group versus control donkeys. (a) There was no significant difference between orthopaedic patient scores and control donkeys with p-value = 0,10 and $N_{patients}$ = 31 and $N_{controls}$ = 35. (b) There was no significant difference between dental patient scores and control donkeys with p-value = 0,39 and $N_{patients}$ = 9 and $N_{controls}$ = 35. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores..

Clinical cases

George Cohen

George Cohen was a 28-year-old gelding presenting with a history of keratoma-like laesion that was confirmed with radiographs, see Fig 14, and a history of osteoarthritis and laminitis. He was on 0,5 gram equipalazone twice a day. The treating veterinarian requested a pain assessment from George, in order to obtain an objective unbiased idea about the amount of chronic pain he was in. He was therefore observed only once by both observers and is not included in the data set of this study. George scored positive on several categories of the FEPSdonkey, including lifted corners of the mouth, abnormal ear position and wrinkles around the eyes, see Fig. 15. George scored 6 out of 24 on the EFPSdonkey pain scale. Furthermore, he presented with a depressed appearance, pressure sores, muscle loss and lameness, see Fig 16. Therefore, he scored 15,5 out of 48 on the Do-CCPS pain scale. This helped the veterinarian and his caretakers to decide to euthanise the donkey.

Figure 14: Clinical case George Cohen, a 28 year old gelding, presenting with keratoma like laesion on radiographs. Dorsopalmar radiograph of the left hindlimb shows a keratoma-like laesion on the lateral side of the hoof, with degradation of the third phalanx.

Figure 15 (left): Clinical case George Cohen, a 28 year old gelding, presented with lifted corners of the mouth, abnormal ear position and wrinkles around the eyes.
Figure 16 (above): Clinical case George Cohen, a 28 year old gelding, presented with pressure sores on his legs.

The donkey was send in for a post mortem inspection. His left hindlimb showed the keratoma like laesion and degeneration of the pedal bone when making horizontal cuts of the hoof, see Fig 17. He also showed signs of cartilage degeneration due to osteoarthritis in his right shoulder joint, see Fig. 18, and laminitic changes such as a thin sole, sinking of the phalanxes and rotation of phalanx 3, see Fig 19. All of which were chronic painful diseases.

Figure 17: Post mortem findings of left hind hoof of clinical case George Cohen when cut horizontally. Arrow shows the keratoma like laesion

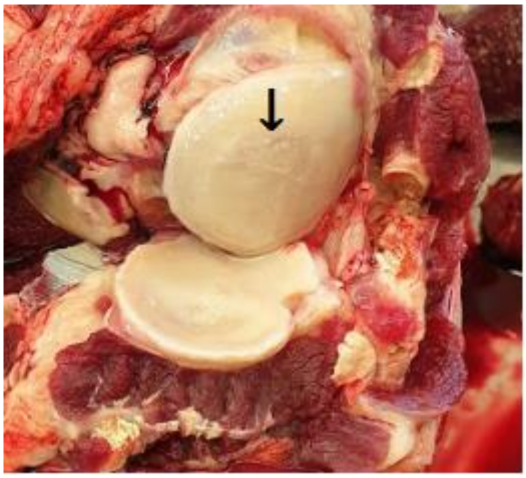


Figure 18: Post mortem findings of right shoulder of clinical case George Cohen. Arrow shows mildly degenerated cartilage.

Figure 19: Post mortem findings of hoof of clinical case George Cohen (left) showing laminitic changes such as a relatively thin sole, sinking of the phalanxes and rotation of phalanx 3. Right photo: hoof of a healthy donkey.

Rosschap Eire

Rosschap Eire was a 9-year-old gelding presenting with a history of osteoarthritis in his right shoulder. Radiographs shown in Fig. 20 were taken over 2 years ago and showed additional bone formation at the right shoulder joint. Rosschap Eire was one of the patient donkeys selected by the veterinarians and was therefore observed six times by both observers. Rosschap got an average FEPS$_{donkey}$ score of 5 out of 24, showing lifted corners of the mouth, abnormal ear position and wrinkles around the eyes and nostrils. Furthermore, he showed severe muscle atrophy of his right shoulder muscles, lameness and positive reactions to flexion tests, resulting in an average of 13,25 out of 48 points on the Do-CCPS scale. Rosschap was due for a re-evaluation by his veterinarian and together with the caretakers, they decided to euthanise the donkey.

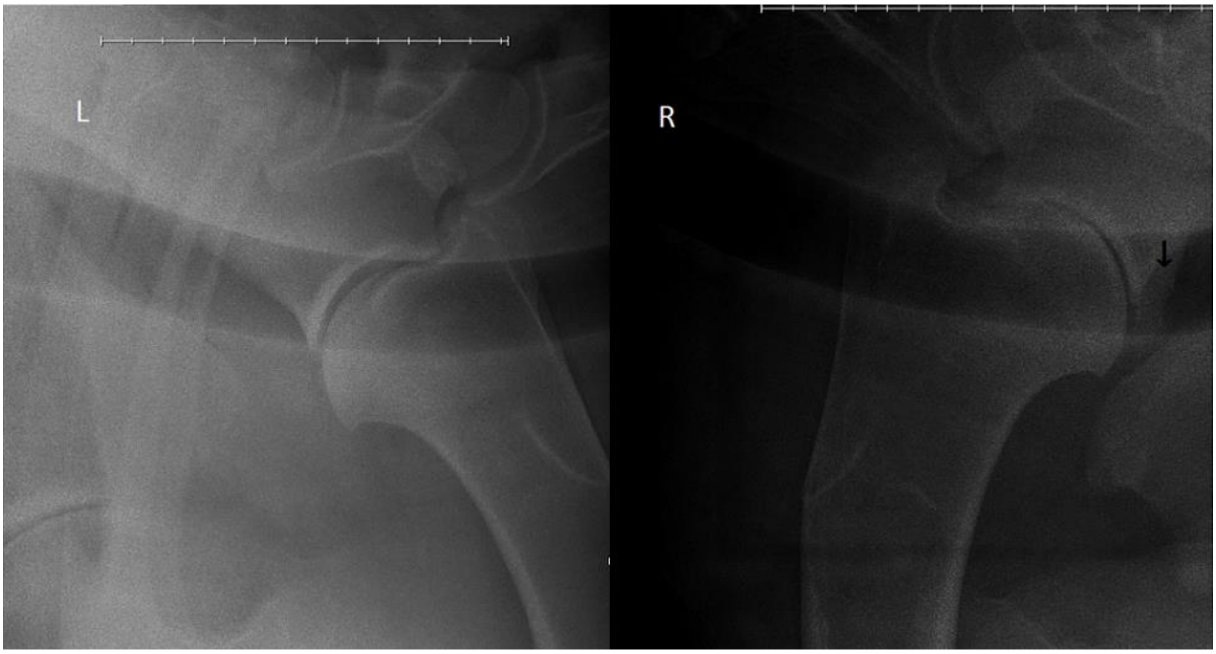


Figure 20: Radiographs of shoulder joints of clinical case Rosschap Eire, a 9 year old gelding. Left shoulder (L) shows healthy radiographs with no osteoarthritic changes. Right shoulder (R) shows severe osteoathrtitic changes with arrow pointing to additional bone formation on the joint ridges. Radiographs were taken 2 years prior to pain assessment of this donkey.

The donkey was send in for a post mortem inspection. His right shoulder showed severe osteoarthritis with cartilage degeneration and additional bone formation on the joint ridges. The joint was almost completely fused together. Severe muscle atrophy was found of the right shoulder musculature. The left shoulder was completely healthy and showed no signs of osteoarthritis, see Fig. 21.

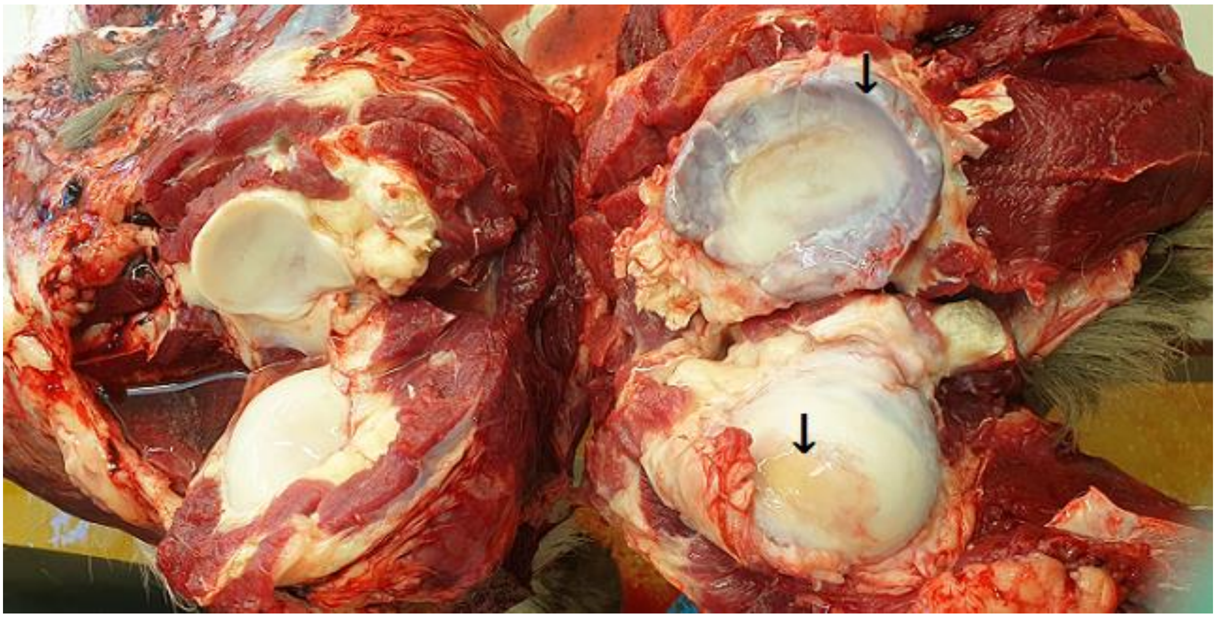


Figure 21: Post mortem findings of clinical case Rosschap Eire, a 9 year old gelding. Left shoulder (left) shows a healthy shoulder joint with no signs of osteoarthritis. Right shoulder (right) shows severe osteoarthritis with cartilage degeneration (lower arrow) and additional bone formation (upper arrow).

Discussion

The research goals for this study were to analyse if the FEPSdonkey scoring system is reliable and clinically applicable for recognising chronic pain in donkeys. This study showed a high interobserver reliability for the FEPSdonkey score between two observers. There was no significant difference found in the patient group over the six time periods. However, the control donkeys showed a significant difference over the six time periods. Therefore, the three average day scores were calculated for each donkey in order to compare patient scores to control scores. For all three days a significant difference was found between patient scores and control scores. Sensitivity and specificity was calculated to be respectively 80,6% and 56,3% for the FEPSdonkey scoring system. Furthermore a survey was developed for the grooms of the donkeys. A significant difference between patient donkeys and control donkeys was found for both the total survey scores and the VAS scores, but also showed a wide variety of scores for the patient donkeys. The total survey score showed no correlation to the VAS score and the FEPSdonkey score. The VAS score only showed a weak correlation to the FEPSdonkey score.

In this study a significant interobserver agreement was found (ICC = 0,96). This means there is a strong correlation between the scores of observer 1 and observer 2. A similar strong correlation was found when using the EQUUS-FAP for horses with acute colic (ICC = 0,93) (van Loon, Van Dierendonck., 2015), head related pain (ICC = 0,92) (van Loon, Van Dierendonck, 2017) and orthopaedic pain (ICC = 0,90) (van Loon, Van Dierendonck, 2019). Other facial expression based pain scores such as the Horse Grimace Scale showed a very good and high correlation when used for assessing horses with acute laminitis (ICC = 0,85) (Dalla Costa, 2016) and routine castration (ICC = 0,92) (Dalla Costa, 2014). Furthermore, FEPSdonkey showed a good correlation when used to assess acute pain in donkeys (R^2 = 0,77) (Van Dierendonck et al., 2018). This is a similar correlation to other studies looking at facial expression changes due to pain in other animal species, for example in humans during a painful procedure (R^2 = 0,71) (Payen, 2001), in sheep due to foot laesions (ICC = 0,86) (McLennan, 2016) and in rabbits due to ear tattooing (ICC = 0,91) (Keating, 2012). The good agreement between the two observers can be explained as an effect of the pain scoring system being as objective as possible. When trying to assess pain in an animal or other person, the Visual Analog Scale can be used. This type of pain assessment is very subjective and previous studies have shown a varying interobserver reliability of poor to moderate, depending on the study (Lindegaard et al., 2010). By making a detailed description of each scoring parameter, the scoring system becomes more objective, which could explain the strong interobserver reliability. Another explanation could be that the strong correlation is the effect of the training the observers did prior to the data collection. The observers did discuss their findings with each other during training and went to the supervisor whenever there was any doubt about scoring one of the categories. The same training method was used prior to collecting the data for validating the FEPSdonkey and Do-CPS for acute pain and resulted in a good correlation for FEPSdonkey and a strong correlation for Do-CPS (Van Dierendonck et al., 2018).

When looking at the pain scores over the 6 time periods, there was no significant difference found for patient donkeys. The patient donkey group showed a stable and higher pain score than the control donkeys. Not all the patient donkeys were on the same medication management, as the different farms handled different medication intervals. The patient donkeys were divided into groups based on their pain management. The pain management groups were 1) receiving Carprofen once a day, 2) receiving Phenylbutazone twice a day at 8 am and 4pm, 3) receiving Phenylbutazone twice a day at 8am and 8pm and 4) receiving no medication. There was no significant difference for each pain management group, however the group that received no medication was very close to being significant (p-value = 0,05 with N = 5). When looking at the data from the simultaneously scored Do-CCPS, this group showed no significant difference (p = 0,38). Based on this data there was no significant difference between the morning and afternoon assessments and between the first, second and third day of assessment for the patient donkeys. However, the group that received no pain relief medication was such a small group, that it is possible that with more data, the difference could become significant. This requires extra attention when doing further research with FEPSdonkey.

Phenylbutazone and Carprofen are both non-steroid anti-inflammatory drugs, but their dosing interval differs. In the Netherlands, phenylbutazone is registered in a dose of twice daily 2,2-4,4 mg/kg bwt, for horses that will not be slaughtered for human consumption. However, phenylbutazone is found to be eliminated much faster in donkeys than in horses. This seems to be the result of donkeys being able to metabolise the phenybutazone faster to the active oxyphenylbutazone and because of a faster clearance of phenylbutazone (170.3 ml/kg/h versus 29.3 ml/kg/h) (Lizarraga, 2004, Mealey et al., 1997). Another study found a clearance of 170-215 mg/kg bwt/h for phenylbutazone in donkeys (Grosenbaugh et al., 2011). This suggests that either the dose for phenylbutazone should be higher when used in donkeys or the dosing interval should be decreased. However, since there are not enough studies done to determine toxicity levels in donkeys, the recommended dose of 2,2-4,4 mg/kg bwt twice daily remains the same for donkeys as it was for horses. In this study most of the donkeys receiving phenylbutazone were on 0,5 g phenylbutazone, regardless of their weight. On average this group of donkeys weighted 194,6 kg (±35,9), resulting in a dose of 2,56 mg/kg. Usually these

donkeys were on what was called ‘quality of life medicine’, meaning they were given this amount of pain medication for the rest of their lives.
The dosing of carprofen for the donkeys in this study varied from 0,61 mg/kg to 1,45 mg/kg. In comparison with phenylbutazone, a much lower clearance of 2,8-4,9 ml/kg bwt/h was found for carprofen. For horses a clearance of 8,4-33,6 ml/kg bwt/h was found (Mealey, 2004). This resulted in a recommended single dose of 0,7 mg/kg bwt for donkeys, similar to the registrated dose for horses in the UK. So even within medication management groups, the donkeys differed in their medication management. It would be useful for further research to see if there is a significant difference over time if all the donkeys in the same management group received the same dose of NSAID’s. Furthermore, carprofen is not registered for horses or donkeys in the Netherlands, but for dogs. Since there are other NSAID’s registered for equines, such as meloxicam, these are the preferred NSAID’s to be used according to cascade.

When looking at the pain scores over the 6 time periods for the patient donkeys, we see a stable higher score over the 6 time periods. However, when looking at the control donkeys, this group did show a significant difference over the six time periods. This means that there might be a variation in the baseline of the pain scores. It suggest that, when using the FEPSdonkey, we should not draw a conclusion based on one observation, but need to do multiple observations over time to determine if the donkey is actually in pain or not. Overall, it was shown that between patient donkeys scores and control donkey scores a significant difference was found, indicating that with repetitive pain scoring patient donkeys would get recognised sooner or later.

The sensitivity and specificity for the FEPSdonkey was calculated to be respectively 80,6% and 56,3%. Furthermore, the individual sensitivity and specificity for each parameter was calculated. 7 parameters turned out to have a sensitivity of close to 0% and specificity of close to 100%, meaning they were rarely or not observed during the observation period. When these parameters were excluded from the data, a slightly higher specificity of 56,7% was calculated, meaning these parameters did not contribute to the ability of the scale to discriminate between patients and controls. Due to limited time, these parameters could get excluded in future studies, while maintaining a good sensitivity of 80,6% and a slightly better specificity of 56,7%. There is however, also reason to keep these parameters included into the scoring system. Parameters such as focus, teeth grinding, moaning and startle might only be seen in severe pain cases and therefore serve as a warning to re-evaluate these donkeys. The EQUUS-FAP has similar scoring parameters as the FEPSdonkey for the head, focus, muscle tone of the head, flehmen and teeth grinding. When the EQUUS-FAP was used to assess horses with acute colic pain, sensitivity for these parameters was respectively 75, 58, 29, 0 and 29%. Specificity was respectively 96, 100, 92, 96, and 100% (van Loon, Van Dierendonck, 2015). This indicates that horses in acute pain do show some of these facial expressions. This could be a reason to keep those parameters included in the FEPSdonkey scoring system. Another solution for the parameters with a low sensitivity and a high specificity could be the use of weighting factors. This would mean that a parameter that has a relatively high sensitivity and specificity would get a greater weighting factor than a parameter that has relatively low sensitivity or specificity. In using a higher weighting factor for the most important parameters, sensitivity and specificity of the overall scoring could improve. Another option could be increasing the dataset with more patients. This could result in an increased sensitivity and/or specificity when more donkeys show these facial expression changes.

The survey that was set up for this study had a return rate of 100% when distributed to the donkeys' grooms. The surveys were handed out to the people who knew the donkeys the best, that being their main caretaker or the farms managers. This meant that one person sometimes filled out multiple surveys for the donkeys on that farm. When looking at the surveys that were filled out at one farm, it came to the attention that none of them had ever scored a zero, not even the control donkeys. The specified orthopaedic and dental questions were filled out with at least a 2, even when the donkeys did not have a history of orthopaedic or dental problems. Furthermore, the VAS score that was given was around 4-5 for each donkey, whether it was a patient donkey or a control donkey. This arose suspicion that the person filling out the surveys at that farm, might have misinterpreted the questions in the survey. Therefore, it was decided not to include the survey results from this particular farm.
Furthermore, when looking at the sensitivity and specificity of each individual survey question, the question about weight loss had a rather low specificity of 61,8%. This question was intended to determine whether or not the donkey was losing weight due to chronic painful problems. For example in horses chronic dental problems could cause the horse to eat less food and chronic orthopaedic problems could cause the horse to walk less in order to get enough food (Taylor, 1997). The same could be true for donkeys in chronic pain. However, there is a possibility that this survey question was not specific enough to be able to get a good answer about this. The farms try to regulate weights of the donkeys very strictly. Donkeys that were too heavy were put on a diet in order to lose weight and donkeys that lost too much weight were put on extra food. All donkeys were weighted monthly to keep track of their weight and would get re-evaluated after. Therefore, it is possible that some donkeys scored positive on the weight loss question for losing weight on purpose. Some additional statistic tests were done to see if the correlation between the survey score and the VAS or FEPSdonkey score would improve if this question was eliminated from the results. Correlation between the survey score and FEPSdonkey score increased slightly from ICC 0,23 to ICC 0,28. Correlation between the survey score and the VAS score decreased from ICC 0,28 to ICC 0,17. Therefore, it was decided to include the results from the question about weight loss in the further statistical tests. In order to get more useful results in the future, this question might need to be reformulated to for example 'did the donkey show unintentional weight loss over time during the last few weeks/months?'.
Another problem presented itself when looking at the last 2 survey questions: in case of chronic lameness/dental problems, the lameness/chewing pattern seen today is normal. These questions would get zero points if the question was filled out with the negative corresponding box and would get 2 or more points if the question was answered with a higher corresponding box. However, later on, it became clear that the higher corresponding box could be interpreted as in worse than normal or better than normal. In order to prevent such confusion, for future research the questions need to be reformulated in: in case of chronic lameness/dental problems, the lameness/chewing pattern seen today is worse than normal. That way, the observers get a better idea of the donkeys overall health over time.
When looking at the reliability and applicability of the survey answers and the VAS score, it showed that the survey total score of patients was significantly higher than the total score of the controls, as did the VAS score for patients versus controls. The VAS score even showed a weak correlation with the FEPSdonkey score of day one (ICC = 0,56). However when looking at the box-and-whisker plots for both these survey and VAS scores, a wide range in scores can

be seen. This again shows the subjectivity of these survey questions and the VAS scores, making it difficult to only rely on survey questions and VAS scores for pain assessment.
Previous studies to chronic pain assessment have reported the use of questionnaires for dog owners. One of these studies reported the use of a descriptive scale of 0, 1, 2, 3, or 4 for answering questions about behaviour and locomotion of dogs diagnosed with canine hip dysplasia, resulting in a chronic pain index. They reported a significant higher chronic pain index in patient dogs versus healthy control dogs. Simultaneously two veterinarians were asked to give a locomotion index score based on lameness, ability to jump and being able to ascend and descend stairs. No correlation was found between the owners chronic pain index and the veterinarians locomotion index. This study suggested that there were some variables that are easier for dog owners to assess and other variables that are easier for veterinarians, suggesting that a combined assessment would be optimal for accurate pain assessment (Hielm-Björkman et al., 2003). Supplementary data 5 shows the results when combining the first 4 questions of this study's survey with the FEPSdonkey day 1 average. A significant difference was found between combined patient donkey scores and control donkey scores, indicating that combining the scores could be an option. However, the box-and-whiskers plot still shows a wide range in scores, indicating there is still much subjectivity in a combination of survey questions and an objective pain score.
Another study reported the use of a descriptive scale of 0, 1, 2, 3, or 4 for answering questions about behaviour and locomotion as well. In this study dogs were assessed that had previously been diagnosed with osteoarthritis in either hip or elbow joints. The dog owners were asked to fill out the questionnaire while one group of dogs were treated with the NSAID carprofen and the other group with a placebo. They found a significant higher pain index in the dogs treated with the placebo, indicating that the dogs in the placebo group were rated as having more pain than the dogs in the carprofen treatment group (Hielm-Björkman et al., 2009). This suggests that the use of questionnaires in dogs can be useful to assess the level of chronic pain and that this could possibly be true for donkeys as well.

When scoring the videos taken for FEPSdonkey scoring during locomotion, a moderate interobserver reliability was found (ICC = 0,69). The pain scoring was done by two different observers in order to be able to blind the observers for the diagnosis and randomise the videos. These observers were familiar with the FEPSdonkey pain scoring system, but one of them was not familiar with pain scoring during locomotion, which proved to be a bit more difficult because not all features of the donkeys face could be seen during the whole video. This could be an explanation for the lower agreement seen now than with the regular FEPSdonkey scoring (ICC = 0,96). The intra-observer reliability was calculated using the 15 doubled videotapes and resulted in a good correlation between the first and second scores (ICC = 0,76). This is a similar intra-observer reliability to a study in horses where an ethogram for pain scoring was used to assess musculoskeletal pain in ridden horses (R^2 = 0,91) (Dyson et al., 2018) and to a study in dogs where the Glasgow pain scale was used to assess postoperative pain in dogs (ICC = 0,85) (Hofmeister, 2018).
Overall, no significant difference was found between patient donkey scores and control donkey scores when using FEPSdonkey during locomotion (p = 0,19), even when looking specifically at the orthopaedic patient group (p = 0,10). An explanation for this is that pain scoring during locomotion on videotape is that not all features could be seen properly on every video, or that with the video editing some pain signals might have been cut out of the 10 second video used for the assessment. Another explanation could be that the donkeys that

were used for this part of the assessment were not all used to walking with a headcollar and rope. The donkeys at the Donkey Sanctuary are all housed in a big group and not regularly asked to walk with the caretakers. The donkeys that would not voluntarily walk while making the video, needed to be persuaded by either pressure on the headcollar or holding a carrot in front of their mouth. This was the case for both patient donkeys and control donkeys and could have influenced some scoring parameters. Possibly, a different outcome would be found if the donkeys used for scoring FEPSdonkey during locomotion were all used to walking with a headcollar and rope. Another possible explanation is that the group of donkeys used for this assessment was not large enough to get a significant difference. It is too early to say that FEPSdonkey can simply not be used for pain assessment during locomotion. One study used the Horse Grimace Scale for pain assessment of horses with laminitis on video (Dalla Costa, 2016). This study found a good inter-observer correlation (ICC = 0,74) and a decreased pain score after treatment. However, the horses in this study were not walking during the videos, making it easier to pain score. Another study investigated the possibility of assessing facial expressions in a group of ridden horses that were lame and a group of ridden horses that were sound and served as a control group (Dyson et al., 2017). They found a significant difference (p<0,001) between lame horses scores and sound horses scores. Overall this study reported changes in ear position, position and expression of the eyes, muscle tension dorsally and caudally of the eye and visibility of the sclera as indicators of facial expression changes due to pain. This study however, took photographs of the horses faces during riding and later assessed them. It indicates that a higher pain score could be seen during locomotion of painful orthopaedic patients, at least when making photographs. Another study was published by the same research group in which they used video footage to assess the facial expressions of horses being ridden. In this study lame horses showed significantly more backwards ear positions, open mouths, tongue out of the mouths, changes in eye posture and expression, head tossing, head tilting and an overall higher pain score than sound horses (Dyson et al., 2018). This indicates that pain assessment in ridden horses is possible, and could be possible for donkeys during locomotion. Further research needs to be done to determine if a difference can be seen between patient donkeys and control donkeys.

While gathering the data for validating FEPSdonkey for chronic pain, the Do-CCPS was also performed (Vos, Unpublished results). The Do-CCPS showed a similarly high interobserver reliability (ICC 0,98) and a significant difference between patient donkey scores and control donkey scores (p<0,001). The Do-CCPS was performed with a cut-off value of 5 and had a sensitivity of 87,8% and specificity of 82,6%. The FEPSdonkey score is also included in the Do-CCPS and contributes greatly to that score. As the Do-CCPS has a higher sensitivity and specificity, it would indicate that using a combination of both pain scales results in a more valid recognition of chronic pain than using the FEPSdonkey alone.

Main limitations of the study

The observers could not be blinded for the donkeys diagnosis prior to pain assessment. This was due to using AMS for looking up age, weight and gender of all the donkeys. Patient donkeys had to be categorised based on their diagnoses and all control donkeys had to be healthy. Therefore all previously diagnosed diseases had to be checked.

The pain score observations were done within the normal routine of the donkeys. This meant that the donkeys were not separated from their group to perform pain assessments, which was good because the observers wanted to observe the donkey in their normal behaviour. However, this also meant that when the observers entered the group, all the curious donkeys gathered around the observers and the donkey that was observed at that time. This might have negatively influenced certain parameters such as ear position when the other donkeys came too close to the observed donkey, resulting in a rather low specificity of that parameter of 62,9%.

The number of donkeys used in the research is not large enough, especially when further dividing the patient donkeys in medication management groups. The small number of patients in each group makes it difficult to perform statistical tests, which will become more reliable with a larger data set.

When making the videotape of the donkeys' faces for the FEPSdonkey scoring during locomotion, some donkeys would not voluntarily walk and had to be persuaded by either pressure on the head collar or by holding a carrot out in front of their mouth. This means that some parameters might have been influenced, because the donkeys were more or less forced to walk instead of walking without being forced.

Future research

For future research it would be helpful to increase the number of donkeys used for pain assessment. It would be interesting to try and keep the observers blinded for the donkeys health status and diagnosis in order not to bias them. If possible, it would be interesting to try to divide patient donkeys into medication management groups in which they receive the same dosing of NSAID's. This is more difficult to obtain because the veterinarians decide the amount of NSAID's the patient donkeys are administered and we would not like to decrease some dosing just for the research, possibly making them more painful. However, with a larger group of donkeys, it might be possible to divide the medication management groups further into dosing groups as well. It would also be interesting to keep making videos of the donkeys' faces for FEPSdonkey scoring during locomotion, to see if a bigger group would change the statistical outcome of the tests. It would also be interesting to see if and how the donkeys would show their facial expressions differently when they are used to walking with a headcollar and rope, ruling out that interference with pain based facial expression changes.

Conclusion

FEPSdonkey shows a very high inter-observer reliability for trained assessors and a significant difference between patient donkey scores and control donkey scores. Because of a variance in the baseline, we should not draw conclusions on one observation, but do multiple observations for each donkey. FEPSdonkey is reliable and applicable for recognising chronic pain in donkeys. The hypotheses 'H0 = the FEPSdonkey is not reliable and applicable for analysing chronic pain in donkeys' is rejected and 'H1 = the FEPSdonkey is reliable and applicable for analysing chronic pain in donkeys' is accepted.

Acknowledgements

A warm and special thanks goes out to the veterinarians from The Donkey Sanctuary for providing the animals for this study and for providing the photos used in the clinical cases. Also special thanks to J.P.A.M. van Loon from Utrecht University for helping with the study design and execution.

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,* vol. 37, no. 6, pp. 565.

Burden, F. & Thiemann, A. 2015, "Donkeys are different", *Journal of equine veterinary science,* vol. 35, no. 5, pp. 376.

Bussieres, G., Jacques, C., Lainay, O., Beauchamp, G., Leblond, A., Desmaizières, L.M., Cuvelliez, S.G. & Troncy, E. 2008, "Development of a composite orthopaedic pain scale in horses", *Research in veterinary science,* vol. 85, no. 2, pp. 294.

Dalla Costa, E. 2016, "Using the horse grimace scale (HGS) to assess pain associated with acute laminitis in horses (Equus caballus)", *Animals,* vol. 6, no. 8, pp. 47.

Dalla Costa, E. 2014, "Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration", *PLoS ONE,* vol. 9, no. 3, pp. e92281.

Dyson, S., Berger, J., Ellis, A.D., Mullard, J. 2017, "Can the presence of musculoskeletal pain be determined from the facial expressions of ridden horses (FEReq)?", *Journal of veterinary behaviour,* vol. 19, pp. 78.

Dyson, S., Berger, J., Ellis, A.D., Mullard, J. 2018, "Development of an ethogram for a pain scoring system in ridden horses and its application to determine the presence of musculoskeletal pain", *Journal of Veterinary Behaviour*, vol. 23, pp 47-57.

Enström A. Objective evaluation of chronic orthopedic pain in a carpal groove model in Shetland ponies. MASTER RESEARCH PROJECT, Utrecht University, Faculty of Veterinary Science, July 2018.

Gleerup, K.B., Forkman, B., Lindegaard, C. & Andersen, P.H. 2015, "An equine pain face", *Veterinary Anaesthesia and Analgesia,* vol. 42, no. 1, pp. 103.

Grosenbaugh, D., Reinemyer, C. & Figueiredo, M. 2011, "Pharmacology and therapeutics in donkeys", *Equine Veterinary Education,* vol. 23, no. 10, pp. 523-530.

Hielm-Björkman, A.K., Kuusela, E., Liman, A., Markkola, A., Saarto, E., Huttunen, P., Leppäluoto, J., Tulamo, R.-. and Raekallio, M., 2003. Evaluation of methods for assessment of pain associated with chronic osteoarthritis in dogs. *Journal of the American Veterinary Medical Association,* 222(11), pp. 1552-1558.

Hielm-Björkman, A.K., Rita, H. and Tulamo, R.-., 2009. Psychometric testing of the Helsinki chronic pain index by completion of a questionnaire in Finnish by owners of dogs with chronic signs of pain caused by osteoarthritis. *American Journal of Veterinary Research,* 70(6), pp. 727-734.

Hofmeister, E.H. 2018, "Agreement among anesthesiologists regarding postoperative pain assessment in dogs", *Veterinary Anaesthesia and Analgesia,* vol. 45, no. 5, pp. 695.

Keating, S.C.J. 2012, "Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses", *PLoS ONE,* vol. 7, no. 9, pp. e44437.

Lindegaard, C., Thomsen, M.H., Larsen, S. & Andersen, P.H. 2010, "Analgesic efficacy of intra-articular morphine in experimentally induced radiocarpal synovitis in horses*",* *Veterinary Anaesthesia and Analgesia,* vol. 37, no. 2, pp. 171-185.

Lizarraga, I. 2004, "Pharmacological and pharmacokinetic differences between donkeys and horses", *Equine Veterinary Education,* vol. 16, no. 2, pp. 102.

McLennan, K.M. 2016, "Development of a facial expression scale using footrot and mastitis as models of pain in sheep", *Applied Animal Behaviour Science,* vol. 176, pp. 19.

Mealey, K.L. 2004, "Pharmacokinetics of R (–) and S (+) carprofen after administration of racemic carprofen in donkeys and horses", *American Journal of Veterinary Research,* vol. 65, no. 11, pp. 1479.

Mealey, K., Matthews, N., Peck, K., Ray, A. & Taylor, T. 1997, "Comparative pharmacokinetics of phenylbutazone and its metabolite oxyphenbutazone in clinically normal horses and donkeys", *American Journal of Veterinary Research,* vol. 58, no. 1, pp. 53-54, 55.

Payen, J.F. 2001, "Assessing pain in critically ill sedated patients by using a behavioral pain scale", *Critical Care Medicine,* vol. 29, no. 12, pp. 2258.

Sutton, G.A., Dahan, R., Turner, D. & Paltiel, O. 2013, "A behaviour-based pain scale for horses with acute colic: scale construction", *The veterinary journal (London, England.1997),* vol. 196, no. 3, pp. 394.

Sutton, G.A., Gila Abells, Paltiel, O., Soffer, M. & Turner, D. 2013, "Validation of two behaviour-based pain scales for horses with acute colic", *The veterinary journal,* vol. 197, no. 3, pp. 646-650.

Taylor, F. 1997, "Investigation of chronic weight loss in adult horses", *In practice,* vol. 19, no. 7, pp. 371.

Van Dierendonck, M.C., Burden, F.A., Richards, K. & van Loon, J P A M 2018, "Objective pain assessment in donkeys – scale construction", *Equine veterinary journal,* vol. 50, pp. 9.

Van Dierendonck, M.C. & van Loon, J P A M 2016, "Monitoring acute equine visceral pain with the equine Utrecht University scale for composite pain assessment (EQUUS-COMPASS) and the equine Utrecht University scale for facial assessment of pain (EQUUS-FAP): a validation study", *The veterinary journal (London, England.1997),* vol. 216, pp. 175.

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 (London, England.1997),* vol. 220, pp. 88.

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 (London, England.1997),* vol. 206, no. 3, pp. 356.

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*, vol. 246, pp 85-91.

Vos, Unpublished results.

Webster, A.J.F. 2001, "Farm animal welfare: the five freedoms and the free market", *The veterinary journal (London, England.1997),* vol. 161, no. 3, pp. 229.

Supplementary Data

Data 1: Differences between the time periods patient donkeys

One of the objectives of scoring twice a day was to determine if there was a difference in the pain scores found in the morning versus in the afternoon. No significant difference was found over the six time periods, with p=0,38, see Fig. 22.

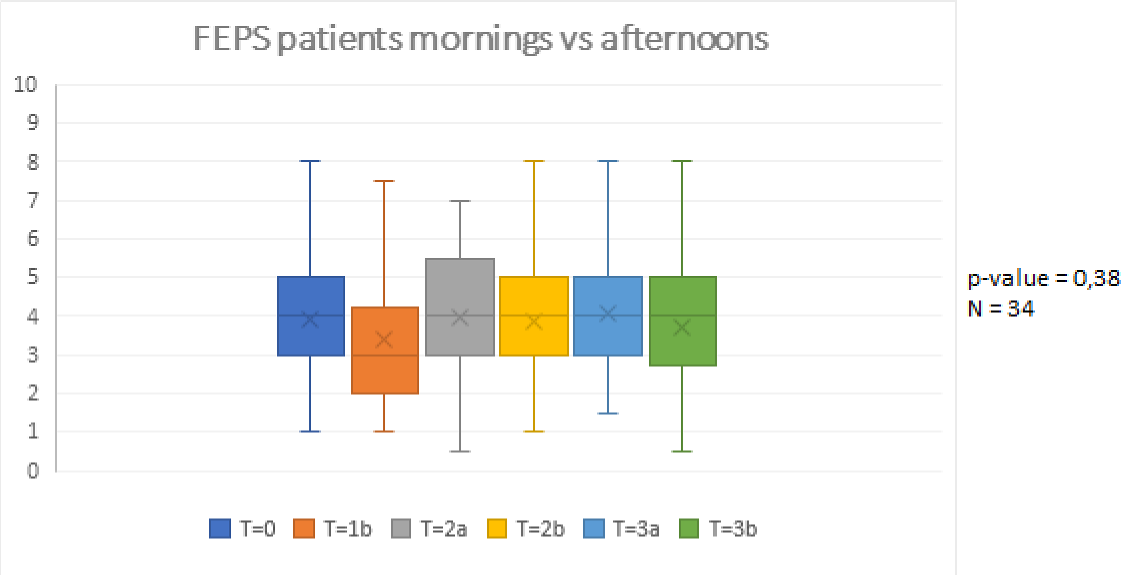


Figure 22: The FEPS scores of the patients over the six time periods. No significant difference was found over the six time periods, with p-value = 0,38 and N=34. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores.

Data 2: Medication intervals

The different farms at the Donkey Sanctuary have different medication regimes. Some farms give out their pain relief medication at 8am and 8pm, others at 8am and 4pm and some at 4pm only. To determine if there was a difference between the morning and afternoon assessment of these donkeys, individual groups of donkeys receiving the same pain medication at the same time were formed.

For the group that was medicated with a single interval dosis (SID) at 4pm, receiving Carprieve® (carprofen), no significant difference was found between morning and afternoon assessment with p=0,76, see Fig. 23. (a).

The group that was medicated with pain relief at 8am and 4pm, receiving Equipalazone® (phenylbutazone), showed no significant difference between morning and afternoon assessment with p=0,99, see Fig. 23. (b).

The group that was medicated with pain relief at 8am and 8pm, receiving Equipalazone® (phenylbutazone), also showed no significant difference between the morning and afternoon assessment with p=0,54, see Fig. 23. (c).

The group of patients that received no pain relief medication at all, showed no significant difference between the morning and afternoon assessment with p=0,05, see Fig. 23. (d).

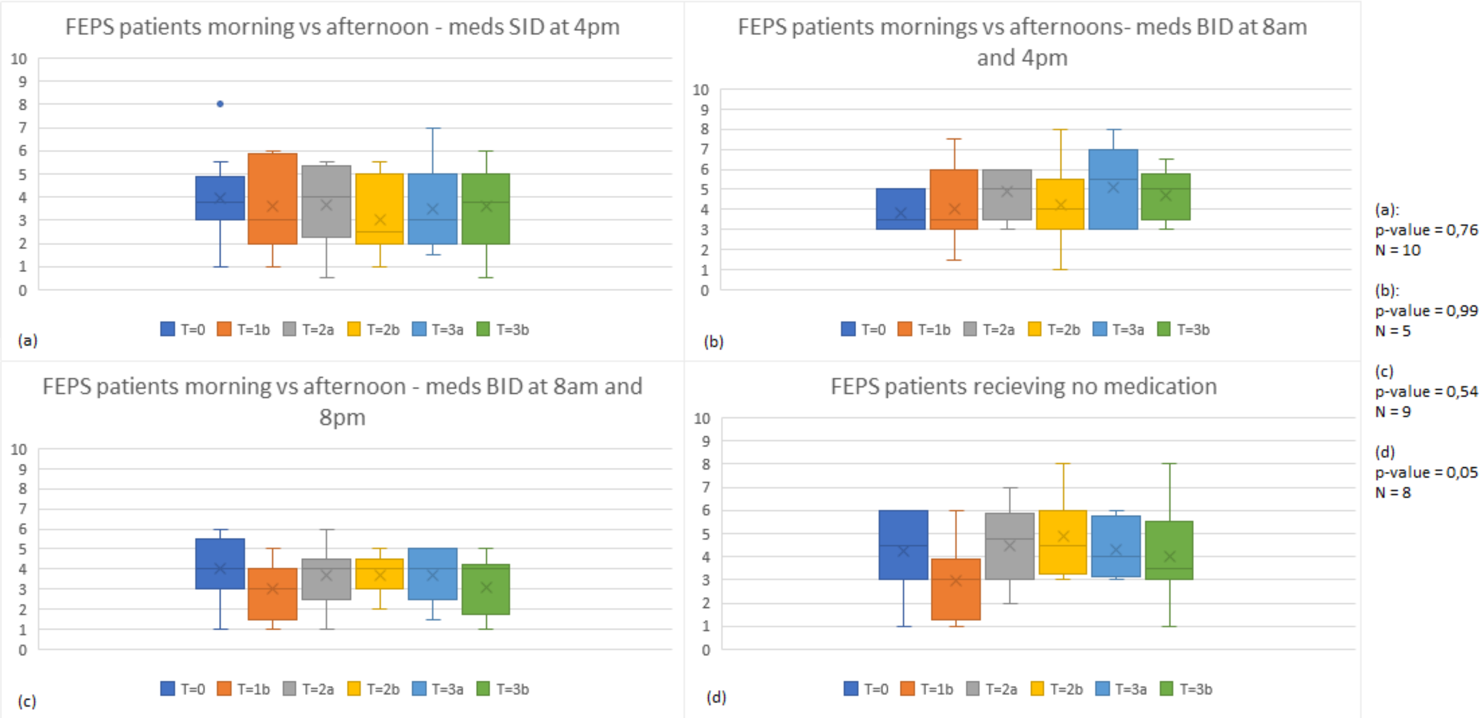


Figure 23: (a) Group of patient donkeys that received their pain relief medication with a single interval doses (SID) at 4pm, showed no significant differences between the morning and afternoon assessment with p=0,76 and N=10. (b) Group of patient donkeys that received their pain relief medication twice a day (BID) at 8am and 4pm, showed no significant differences between the morning and afternoon assessment with p=0,99 and N=5. (c) Group of patient donkeys that received their pain relief medication twice a day (BID) at 8am and 8pm, showed no significant differences between morning and afternoon assessment with p=0,54 and N=9. (d) Group of patient donkeys that received no medication showed no significant differences between morning and afternoon assessment with p=0,05 and N=8. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.

Data 3: Differences between the time periods control donkeys

The same test was also performed for the control donkeys, to determine if there was a significant difference over the six time periods, see Fig. 24. For the controls, a significant difference was found over the six time periods, with p<0,01.

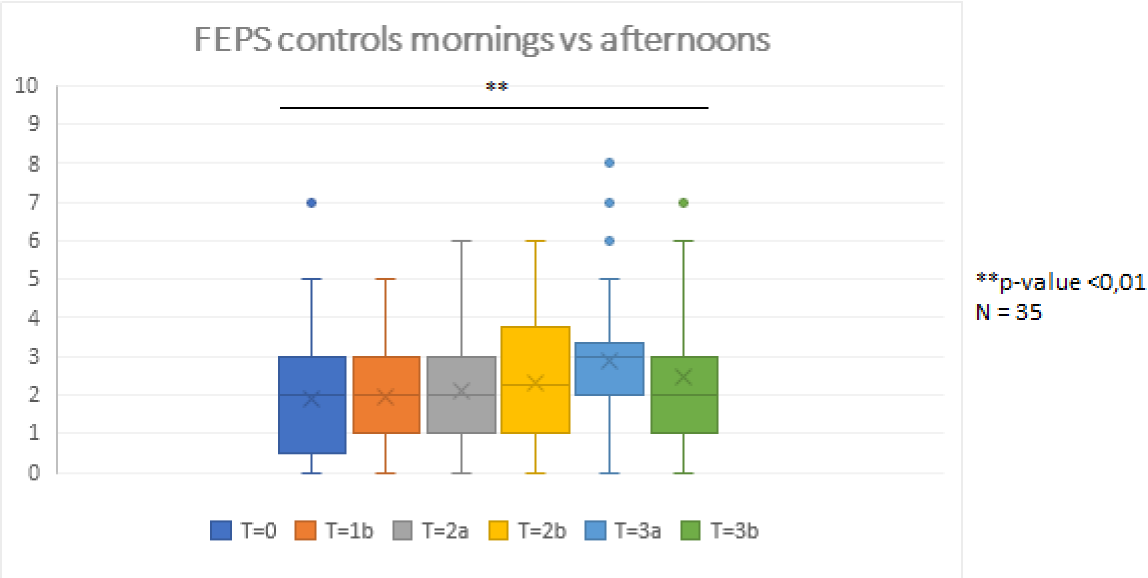


*Figure 24: The FEPS scores of the controls over the six time periods. A significant difference was found over the six time periods, with **p-value <0,01 and N=35. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.*

To determine if there was a difference between the morning and afternoon assessment of respectively day one, two or three, some further testing was done, see Fig. 25. No significant difference was found between the morning and afternoon assessments of day one, two or three, with p-values of respectively p= 0,88, p= 0,60 and p= 0,13.

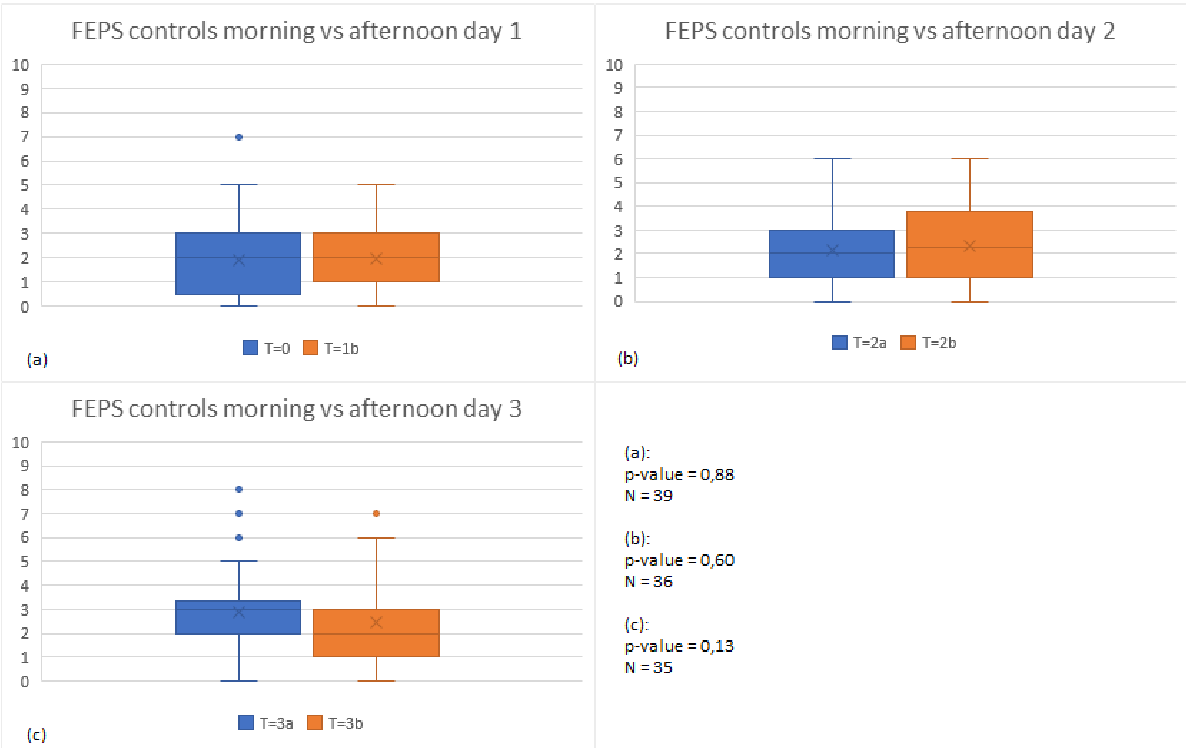


Figure 25: No significant difference was found between the morning and afternoon assessments of respectively day one, two and three of the control donkeys. (a): p-value = 0,88 with N=39; (b): p-value = 0,60 with N = 36 and (c): p-value = 0,13 with N = 35. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.

Data 4: Survey data results filled out about patient donkeys (Table 6) and control donkeys (Table 7)

Table 6: Survey data patient donkeys. Yellow rows show the one farm that possibly interpreted the questions differently. Of the five donkeys, only one donkey (Peg E Eire) was a dental and orthopaedic patient and therefore might have both specific orthopaedic and dental questions filled out positive. Results from this farm were excluded from the data.

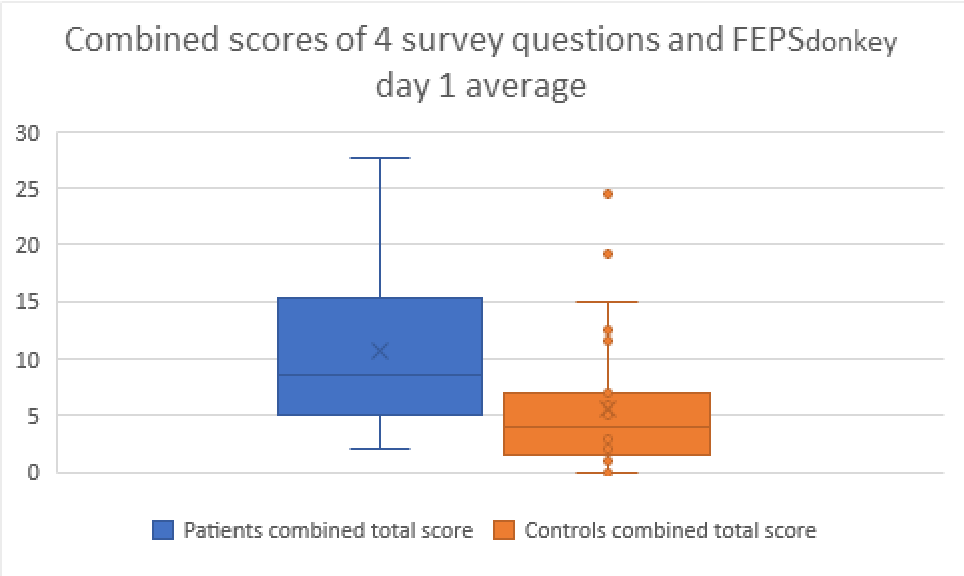
Name patient donkey	Mindset is alert	Weight loss	Interacts less	Lays down more	Chronic lameness	Chronic dental	Survey total score patients	VAS score patients
Jacob Eire	0	4	0	0	0	0	4	4
Ringo Retford	0	10	4	0	0	0	14	3
Darcy Eire	4	2	8	8	2	0	24	4
Fergie T	0	8	2	0	0	2	12	4
Brandy Hubbard	2	8	0	0	0	0	10	8
Twiggy Hubbard	0	10	0	2	0	0	12	7
Leia Molly	0	8	0	0	0	0	8	7
Gem Hubbard	0	0	0	0	0	0	0	9
Chrysanthamum Swain	2	2	2	2	4	2	14	5
Catkin Curry	2	6	4	6	2	2	22	6
Hettie O'Keefe	2	2	6	2	4	2	18	4
Peg E Eire	4	4	4	0	2	2	16	4
Sheena McNally	2	2	2	2	2	2	12	4
Harry Jeffords	0	10	0	0	0	0	10	3
Nell Morgan	2	10	10	0	0	0	22	5
Timmy Wilson	4	0	4	0	0	0	8	3
Gwyneth BV	2	10	6	0	4	0	22	3
Rosschap Eire	0	0	0	0	2	0	2	6
DD	0	0	0	0	0	0	0	5
Bella Cook	0	0	0	0	0	0	0	4
Truffel Stone	0	2	0	0	0	0	2	5
Willy F	0	2	0	0	0	0	2	4
Cinnamon Bridgen	0	0	0	0	0	0	0	4
Bump Eire	2	10	6	0	0	0	18	7
Peanuts Williams	0	4	0	0	0	0	4	0
Portnod Lass	0	4	4	0	0	0	8	1
Ted Assenheim	0	0	10	0	0	0	10	0
Ben Sargent	0	0	0	0	0	0	0	0
Daffy H	0	4	0	0	0	0	4	2
Malty Eire	0	2	2	0	0	0	4	2
Barney ILPH	0	0	2	0	0	0	2	0
Ella Webber	0	0	0	0	0	0	0	1
Kerry II Eire	0	0	6	0	0	0	6	1
Muffin Simmons	0	0	0	0	0	0	0	0
Gavroche	2	2	4	2	2	2	14	1
Barton	2	2	2	2	2	4	14	1
Barney Conti	4	2	6	2	2	2	18	5
Mr Pickwick	0	0	0	0	2	0	2	1

Table 7: Survey data control donkeys. Yellow rows show the one farm that possibly interpreted the questions differently. All five control donkeys from that farm were healthy and had no history of orthopaedic or dental problems and therefore both specific orthopaedic and dental questions should have been filled out negative. Results from this farm were excluded from the data.

Name control donkey	Mindset is alert	Weight loss	Interacts less	Lays down more	Chronic lameness	Chronic dental	Survey total score controls	VAS score controls
Selwyn	0	0	0	0	0	0	0	2
Ken Eire	0	0	0	0	0	0	0	2
Archie Keevans	0	0	0	0	0	0	0	0
Floss Morris	0	2	2	0	0	0	4	1
Huwi	0	0	0	0	0	0	0	0
Bluebell RSPCA	0	0	0	0	0	0	0	0
Toby AT	0	0	0	0	0	0	0	0
Tufty AT	4	0	6	0	0	0	10	0
Norman Eire	2	0	2	0	0	0	4	0
Nina Ball	2	2	4	4	2	2	16	4
Jenny Mumsy Tinney	2	2	2	2	2	2	12	5
Daisy Davidson	2	2	2	2	4	2	14	3
Barbara Douglas	4	2	6	2	2	2	18	4
Caitlin E	2	2	2	2	2	2	12	5
Harry Howe	0	10	8	0	0	0	18	0
Dylan Wilson	0	4	6	0	0	0	10	0
Eastern Martin	6	10	4	2	6	0	28	0
Chicoe Eire	0	0	0	0	0	0	0	0
Jacko Mc Eire	0	0	0	0	0	0	0	0
Squirrel Cameron	0	6	0	0	0	0	6	0
Princess Stone	0	0	0	0	0	0	0	0
Jenny Johnson	0	2	0	0	0	0	2	0
Poppy Roebuck	0	0	0	0	0	0	0	0
Neddy Woods	0	4	0	0	0	0	4	0
Kate Morgan	2	4	2	0	0	0	8	0
Lucky Williams	0	0	0	0	0	0	0	0
Isobel ST	0	0	0	0	0	0	0	0
Bee Jay Conti	0	0	0	0	0	0	0	0
Barney Steel	0	6	0	0	0	0	6	0
Freddie Cole	2	2	4	4	0	0	12	1
Eeaaw Eire	0	6	0	0	0	0	6	0
Bessie Cusack	0	0	0	0	0	0	0	0
Archie Bridger	0	0	0	0	0	0	0	0
Jack F Eire	0	0	2	0	0	0	2	0
Timmy Wood	0	0	0	0	0	0	0	0
Jacko Buttle	0	0	4	0	2	2	8	2
Mr Darcy Atkinson Eire	0	0	0	0	2	4	6	1
Mary Sanderson	0	2	0	0	2	0	4	0
Timothy Baker	0	2	0	0	0	0	2	0

Data 5: combining the survey results with FEPSdonkey

The survey results were be combined with the FEPSdonkey results to determine if there was a significant difference between patient and control donkey scores. In order to combine the result from the survey and FEPSdonkey score, the results of the first 4 questions of the survey and the FEPSdonkey day 1 average score were added. Patient donkey scores and control donkey scores can be seen in Fig. 26. A significant difference was found between combined patient donkey scores and combined control donkey scores with p-value <0,001.



***p-value <0,001
Npatients = 33
Ncontrols = 34

*Figure 26: When combining the results of the first 4 questions of the survey with the FEPSdonkey day 1 average scores, a significant difference was found between patient donkey scores and control donkey scores with ***p-value<0,001 with Npatients = 33 and Ncontrols = 34. Boxes show the 25-75th percentiles, lines in the boxes show the median scores, crosses in boxes show the average scores and the end of the whiskers show minimum and maximum scores. Dots show the outliners.*